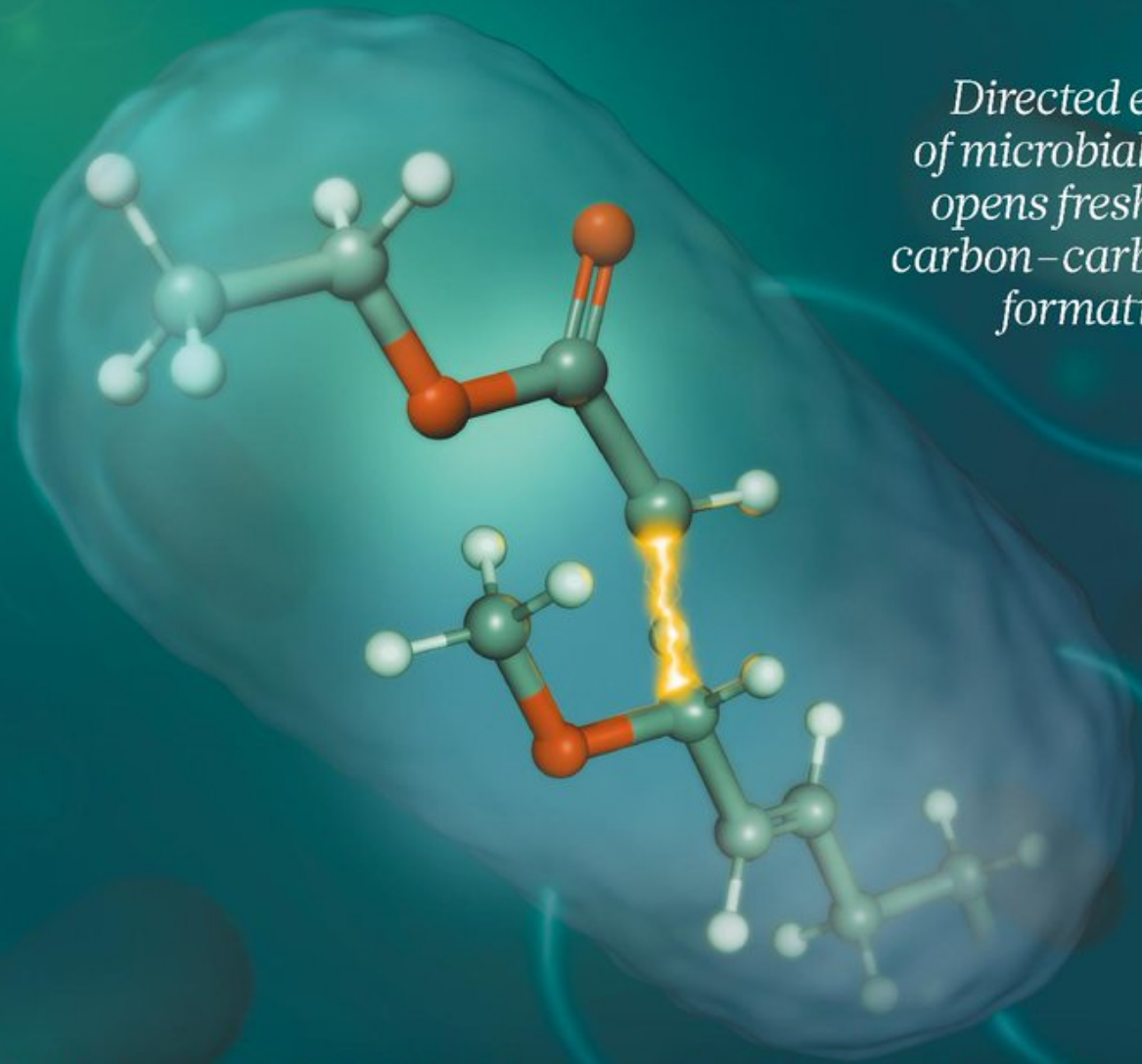
nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE



Directed evolution of microbial enzyme opens fresh route to carbon–carbon bond formation **PAGE 67**

UNNATURAL SELECTION

CONDENSED-MATTER PHYSICS

MAGIC ANGLE

Graphene puts a fresh twist on superconductivity

PAGE 15

CULTURE

THE HOT TICKETS IN 2019

Must-see exhibitions, plays and films

PAGE 22

BIOCHEMISTRY

KIDNEY PROTECTION

Nitric oxide acts to save renal function during injury

PAGES 33 & 96

NATURE.COM

3 January 2019

Vol. 565, No. 7737

THIS WEEK

EDITORIALS

GREEN PLAN Deal hinges on stopping subsidies to polluters p.6



SPACE ROCK NASA craft flies past object 6.5 billion kilometres away p.7

WHALING Outcry as Japan withdraws from IWC to resume hunting p.7

Nature at 150

In the 150 years since the first issue was published, Nature *has evolved alongside the research community it serves. We hope to continue to grow in the years to come.*

In 1869, the US state of Wyoming passed the world's first law that enshrined women's right to vote. Leo Tolstoy published the final volumes of his epic novel *War and Peace* and the Suez Canal in Egypt was opened. Mahatma Gandhi was born. And the University of Oxford competed against Harvard University in the first international boat race, held on London's River Thames.

The world was a different place in those days, but, 150 years later, some things remain the same. *War and Peace* continues to be enjoyed by many, and Wyoming policymakers proved to be welcome pioneers of a much broader and more influential cause. Another constant is the journal you are currently reading. For 1869 also saw the publication of the first issue of *Nature*. That makes this year our sesquicentennial. And that's a cause for celebration, reflection and gratitude — to the global research community and its evolving needs that have helped to shape and guide us, down through the decades, into what we are today.

Nature wasn't the first and isn't the oldest scientific journal. But celebrate we shall — mostly in November, which will mark 150 years since the official start of our weekly issue.

To look at the history of *Nature* is to trace how science, the political context within which it operates, and its communication have evolved over the past 150 years. And so, during the course of the year, we will gradually examine the progress of scientific endeavour across many disciplines, and consider how science's role in broader society has changed.

We will explore the legacy of some of the most influential research papers we have published over the years. And we will delve into our archive for the most interesting of our historical content and try to put it into a broader context. We also intend to share the anniversary with readers: over the course of the year we will be inviting you to contribute your thoughts on the future of research and its dissemination.

Anniversaries are first and foremost an opportunity to reflect. Looking back on the science of 1869, we discover that it was the year when Friedrich Miescher isolated what he called nuclein (now known as nucleic acids), from the nuclei of white blood cells. That same year, Paul Langerhans first described the pancreatic islets, Dmitri Mendeleev first presented the periodic table and Alfred Russel Wallace published *The Malay Archipelago*, in which he described the division of fauna and flora along what is now known as the Wallace line.

The name of our journal has remained the same throughout the past century and a half. But much of what lies between the covers has changed — with the times, with evolving science and with publishing technology.

In 1869, scientific journals served mainly to record presentations made at meetings of learned societies or to reprint valuable papers published elsewhere, perhaps in another language. At the end of the nineteenth century, the learned societies, associations and institutions were more prominent than journals in shaping the progress of science. In fact, it was not until the Second World War that scientific journals became the main forum for disseminating primary research findings.

When *Nature* was launched, it was intended to be more like the *Scientific American* or *New Scientist* of today — its first editor, Norman Lockyer, wanted "men of science" to write about their work for the general audience. (Although clearly there were also women of science at the time, they went unacknowledged — as they did in so many walks of life.)

"The name of our journal has remained the same, but much of what lies between the covers has changed."

Despite Lockyer's intentions, *Nature* refocused on serving the professional research community within just a few years of its launch, when early-career researchers of the time realized they could use the journal and its rapid weekly publication cycle to their own advantage — to advance scientific discourse in the community and, eventually, to publish their original findings.

At the time of *Nature*'s launch, the predominant mode of communication between scientists was through personal letters, and a formal way of disseminating work involved publishing monographs. One prominent example from the era immediately springs to mind — Charles Darwin's *On the Origin of Species*, which was published in 1859.

Nature is known for its original research as well as its news journalism and commentary. Today this is spread across different formats, but the goal of the coverage remains the same: to help readers make sense of the world of science; to help them in their work and in their careers; and to help them to assess the position of science in the context of society. As such, *Nature* has traditionally weighed in on broader political and societal issues — in editorials and elsewhere. And we will continue to do so.

Nature in its early days focused predominantly on science done in Britain. As today's research is a global endeavour, so is our focus, both in the original research we publish and in our news coverage. The global nature of science today necessarily means that it has become more collaborative. And, as we have gone from publishing papers with one or just a few authors to papers authored by large consortia, we have worked to acknowledge the contributions of individual authors.

Looking forwards is also an important aspect of any anniversary celebration. We will do so throughout the year, in part for our own benefit, to consider how we can best continue to evolve with the research community and its needs, striving to build on our efforts to support reproducibility, diversity and social justice in research. We hope and expect 2019 to be another significant year in science. And we look forward to sharing it with you. ■

150 nature
150 YEARS OF NATURE
Anniversary collection
go.nature.com/nature150

© 2019 Springer Nature Limited. All rights reserved.



COLORADO STATE UNIVERSITY

How to make the next Green New Deal work

To make green investments pay off, policymakers must learn from past mistakes and stop subsidizing polluters, urges **Edward B. Barbier**.

As the 116th US Congress begins, a coalition is growing around an ambitious Green New Deal. If successful, a new House of Representatives committee would craft a 10-year plan to shift away from polluting industries, embrace green infrastructure and produce 100% of energy from renewables, improving prospects for US workers.

Sound familiar? It is. In 2008, in the midst of the Great Recession, the United Nations Environment Programme asked me to write a report that formed the basis of its Global Green New Deal to stimulate economic recovery and create jobs. It aimed to improve the lives of the world's poor, lessen carbon dependency and reverse environmental degradation.

In the decade since, I have watched what worked, what didn't and why. For the latest Green New Deal to flourish, the US government must first end fossil-fuel subsidies and correct other market distortions that prop up 'brown economies' — those that rely on fossil fuels and ignore the environmental impacts. Second, it must finance the new policy sustainably.

There were high hopes for the UN's Global Green New Deal. Between 2008 and 2010, the G20 nations and a handful of other economies put US$3.3 trillion into fiscal stimulus, of which more than $520 billion was devoted to 'green investments'. This included pollution clean-up, recycling and low-carbon energy. More than 60% of the green stimulus went to improving energy efficiency, with an aim to create much-needed jobs in construction (E. B. Barbier *Can. Public Policy* **42** (Suppl. 1), S1–S9; 2016).

China invested 3% of its gross domestic product (GDP) and South Korea put in 5% of GDP as part of long-term strategies to develop industries around such technologies as solar panels, electric cars and wind turbines. Other economies spent much less: the United States devoted about 0.9% of GDP, with Canada and the European Union investing around 0.2%. Since the global economic recovery began in 2010, there has been scant additional support for this green transition.

Meanwhile, China has become the leading producer of solar cells, wind turbines, energy-saving lights and solar water heaters. It aims to be the market leader in fuel-efficient cars. South Korea has also expanded exports from green industries, including an ambitious plan implemented during 2009–13 to create 1.6 million to 1.8 million jobs through green growth by 2020. In most major economies, however, green sectors have largely been left to develop on their own, and remain niche. For example, in the United States, sectors such as renewable energy, pollution abatement, materials recycling and conservation employ just over three million workers and account for 3% of GDP.

The brown economy remains pervasive, partly because it is buttressed by market-distorting subsidies. The International Monetary Fund (IMF) estimated the global distortion for fossil-fuel subsidies alone at $5.3 trillion in 2015, or 6.5% of GDP (D. Coady *et al. World Dev.* **91**, 11–27; 2017). Subsidies for agriculture, water and transportation also reward polluting activities and the overuse of resources.

Federal and state governments should eliminate harmful subsidies and use pollution taxes and carbon pricing to account for the toll on human health and on natural capital (clean air, functioning ecosystems, and so on). Fees, tradeable permits and other market mechanisms would put a cost on pollution, carbon emissions and excessive resource use. Such green incentives are doubly productive. They benefit health and the environment, and stimulate sustainable growth.

By the end of 2018, 46 countries and 25 sub-national jurisdictions were pricing carbon, accounting for around 20% of global greenhouse-gas emissions and raising $82 billion in revenue (see go.nature.com/2bvften). The IMF estimated that removing fossil-fuel pricing distortions would cut global carbon emissions by 21%, reduce deaths from fuel-related air pollution by 55% and raise extra revenue of 4% of global GDP in 2013.

REVENUES FROM DISMANTLED SUBSIDIES CAN BE PUT TOWARDS A SUSTAINABLE AND EQUITABLE FUTURE.

Such revenues should be used to finance government investments in greening the economy. These funds could support better infrastructure for renewable energy, more sustainable urban development and research into clean energy. They could also be used to raise the minimum wage, provide payments or retraining for displaced workers, and reduce burdens for vulnerable households affected by the green transition. In short, revenues that come from dismantled subsidies and environmental taxes can be put towards a sustainable and equitable future.

The Green New Deal should not be funded with deficit spending. Saddling future generations with unsustainable levels of national debt is just as dangerous as burdening them with an economy that is environmentally unsustainable. Deficit spending is warranted to boost overall demand for goods and services when unemployment rises, consumers do not spend and private investment is down. When that is not the case, efforts to boost green sectors should pay for themselves.

Crafting a successful Green New Deal will be hard work. In 2009, the G20 promised to phase out fossil-fuel consumption subsidies, but so far only Indonesia has implemented substantial reforms. Taxes and carbon pricing have always faced stiff political resistance, especially in the United States. Most economies have a poor track record of long-term planning, which will be needed for any green investment strategy. And yet, the US Green New Deal represents the first time a major Western economy has proposed a comprehensive ten-year plan for a green transition.

Proponents are correct that we urgently need a strategy to build a sustainable economy. Let's get it right this time. ■

Edward B. Barbier *is an economics professor at Colorado State University in Fort Collins and author of* A Global Green New Deal (Cambridge Univ. Press, 2010).
e-mail: edward.barbier@colostate.edu

© 2019 Springer Nature Limited. All rights reserved.

SEVEN DAYS *The news in brief*

EVENTS

Indonesian tsunami

Part of Indonesia's Anak Krakatau volcano collapsed on 22 December during an eruption, triggering a tsunami that hit the coasts of Java and Sumatra and killed at least 430 people. Satellite and aerial images later confirmed that much of the western flank of the volcano had disappeared into the sea. The blast reduced Anak Krakatau to one-third of its previous height. The volcano had been erupting since June. It originally formed from the remains of the volcano Krakatau, whose 1883 eruption generated a tsunami that killed at least 36,000 people.

SPACE

Ultima Thule fly-by

On 1 January, NASA's New Horizons spacecraft flew past the space rock 2014 MU_{69}, nearly 6.5 billion kilometres from Earth — the most distant Solar System object ever visited. Early images, taken before the spacecraft whizzed just 3,500 kilometres above MU_{69}'s surface, showed an elongated blob that resembles a bowling pin. The object — which researchers have nicknamed Ultima Thule — is spinning almost directly face on to Earth, like a propeller blade.

POLICY

Japanese whaling

Japan has been condemned for its decision to withdraw from the International Whaling Commission (IWC) and resume commercial whaling. The government announced on 26 December that it will begin commercially hunting the mammals in its waters this year, but will end its whaling programme in the Southern Ocean. The IWC, based in Cambridge, UK, introduced a moratorium on whaling in 1986, but Japan has continued to hunt the mammals, citing scientific purposes (**pictured**, minke whale at Kushiro Port). Last year, the IWC rejected the government's bid to restart commercial whaling. Japan says it will now participate in the IWC only as an observer.



THE ASAHI SHIMBUN/GETTY

POLITICS

US shutdown

Several major US science agencies shut down indefinitely on 22 December after politicians failed to reach a deal to continue funding government operations. NASA, the National Science Foundation, the Environmental Protection Agency and the Food and Drug Administration are among the agencies affected by the shutdown — the third in 2018. By law, they must curtail all activities except those considered essential for protecting life and property. It is not clear when the funding impasse will end; President Donald Trump says that any budget deal must include US$5 billion to construct a wall along the US border with Mexico, but Democrats in Congress say they will not vote for such a plan.

Ebola troubles

Political protests are thwarting Ebola control efforts in Beni and Butembo, northeastern cities in the Democratic Republic of the Congo. On 28 December, the World Health Organization reported that it is struggling to carry out measures such as identifying people potentially infected with the virus. In Beni, protesters robbed and set fire to an Ebola centre. The unrest follows President Joseph Kabila's decision to suspend voting in presidential elections in Beni, Butembo and Yumbi, which are opposition strongholds. Observers suspect that his motivation is to suppress votes.

TREND WATCH

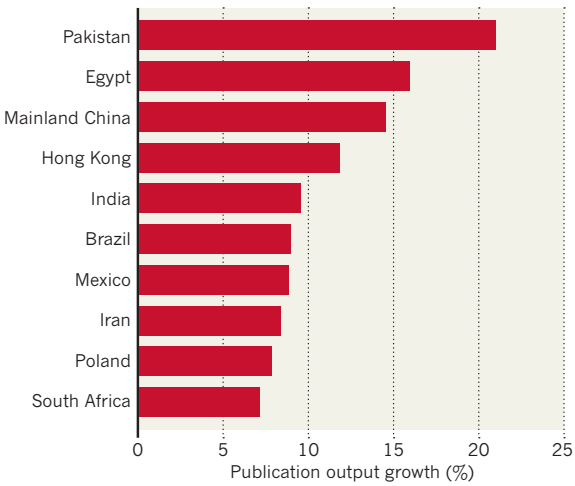
Emerging and developing economies showed the largest increases in research output in 2018. Pakistan and Egypt topped the list in percentage terms, with rises of 21% and 15.9%, respectively. China's publications rose by about 15%, with India, Brazil, Mexico and Iran all seeing their output grow by more than 8% compared with 2017.

This diversification of players in science is a phenomenal success, says Caroline Wagner, a science-policy analyst at the Ohio State University in Columbus. "In 1980, only 5 countries did 90% of all science — the United States, the United Kingdom, France, Germany and Japan," she says. "And now there are 20 countries within the top producing group."

Globally, the output rose by around 5% in 2018, to an estimated 1,620,730 papers listed in the Web of Science science-citation database. The data were compiled for *Nature* by Clarivate Analytics, owner of Web of Science, which says the overall rise is comparable to increases over the past few years, and is likely to continue in 2019. It's not yet clear what is driving the rises in emerging nations; they could be partly due to changes in how the database is curated, to add more local and national journals.

COUNTRIES WITH BIGGEST RISES IN RESEARCH OUTPUT

Emerging economies top the list for percentage increase in publications from 2017 to 2018.



SOURCE: INSTITUTE FOR SCIENTIFIC INFORMATION, CLARIVATE ANALYTICS

© 2019 Springer Nature Limited. All rights reserved.

NEWS IN FOCUS

PHYSICS Tantalizing signs of superconductivity at near-room temperature **p.12**

POLITICS Violence in Nicaragua engulfs scientists **p.11**

NEW YEAR Gene-editing, open access and seals with sensors to shape 2019 **p.13**

MATERIALS The scramble to understand a twisted form of graphene **p.15**



THE ASAHI SHIMBUN VIA GETTY

Japan's Kamioka Gravitational Wave Detector is scheduled to start up in 2019, joining a global network of interferometers.

PHYSICS

Japan to begin pioneering hunt for gravitational waves

The underground KAGRA detector will deploy ambitious technology to improve sensitivity.

BY DAVIDE CASTELVECCHI

Inside a house-sized scaffolding wrapped in thick plastic sheets, Takayuki Tomaru is in full clean-room attire. The physicist, who works at the High Energy Accelerator Research Organization (KEK) in Tsukuba, Japan, is performing one of the most delicate and crucial tasks in the construction of a gravitational-wave observatory: installing one of the machine's four mirrors, each a 23-kilogram cylinder of solid sapphire known as a test mass. When operations begin later this year, their job will be to bounce infrared laser beams back and forth along two 3-kilometre, high-vacuum pipes, ready to sense the passage of gravitational waves (see 'Japan's wave hunter').

The ¥16.4-billion (US$148-million) observatory — Japan's Kamioka Gravitational Wave Detector (KAGRA) — will work on the same principle as the two detectors of the Laser Interferometer Gravitational-Wave Observatory (LIGO) in the United States and the Virgo solo machine in Italy. In the past few years, these machines have begun to detect gravitational waves — long-sought ripples in the fabric of space-time, created by cataclysmic cosmic events such as the merging of two black holes or the collision of two neutron stars.

With the addition of KAGRA, the growing global network of detectors will enable astrophysicists to locate the position of these feeble cosmic signals in the sky with greatly increased precision. They will be able to dissect the waves' properties, such as how they are ▶

© 2019 Springer Nature Limited. All rights reserved.

oriented in space, better than ever before, ultimately revealing more about the elusive cosmic objects that produce them.

But KAGRA will not just be more of the same — it could prove to be an important test bed for future detectors. "KAGRA is carrying out tests of two concepts that may prove important for the future of gravitational-wave astronomy," says Rainer Weiss, a physicist at the Massachusetts Institute of Technology in Cambridge who co-founded LIGO.

One innovation is that it is the first major interferometer to be built underground. Its two arms stretch inside tunnels under Mount Ikenoyama, near Japan's north coast. "We believe this is an advantage, because seismic noises are typically two orders of magnitude smaller underground," says Takaaki Kajita, a physicist at the University of Tokyo who is KAGRA's principal investigator.

And whereas LIGO's and Virgo's mirrors operate at room temperature, KAGRA's will be kept very cold, at 20 kelvin, to cut down noise from thermal vibrations.

If KAGRA works as planned, it could provide crucial know-how for the field. The use of cryogenics, in particular, might be essential if future detectors are to offer vastly improved sensitivity, says physicist David Shoemaker at the Massachusetts Institute of Technology, who is LIGO's spokesperson.

A WORK IN PROGRESS

Japan was an early starter in the race to detect gravitational waves, the existence of which Albert Einstein predicted more than a century ago. Researchers at the University of Tokyo built prototype interferometers in the early 1990s, following work by physicists in the United States, the United Kingdom and Germany. When TAMA, a machine with 300-metre arms, began operating in 1998, it was the largest, most sensitive prototype detector in the world, says physicist Raffaele Flaminio, who has for several years been a leading KAGRA researcher at the National Astronomical Observatory of Japan. But TAMA wasn't expected to make a discovery. Because gravitational waves stretch the dimensions of space, the waves' effects are most noticeable over long distances, putting smaller detectors at a disadvantage. Moreover, human-generated vibrations made TAMA — located in Tokyo — a non-starter, Kajita explains.

In the 1990s, researchers in Europe and the United States secured funding to build LIGO, two 4-kilometre interferometers in Washington state and Louisiana, and the 3-kilometre Virgo, but Japanese physicists faced an uphill struggle for funding. A further setback came in 2001, when a serious and costly accident at Super-Kamiokande, a massive neutrino observatory also under Mount Ikenoyama, made the Japanese government wary of funding big-science projects.

"It's a daunting but courageous thing that KAGRA has done."

Still, Japanese researchers kept working on interferometer development, pursuing the cold-mirror route. In 2006, the Cryogenic Laser Interferometer Observatory (CLIO) began operating in a tunnel at Kamioka. The 100-metre prototype was the first to have cryogenically cooled mirrors, and took two decades to perfect, Kajita says. That's in large part because cryogenic cooling poses a conundrum for gravitational-wave science. "Coolers are mechanical things," Kajita says, so they create vibrations of their own. Researchers had to work out how to keep the coolers in physical contact with the mirrors' suspensions — so that they could keep the mirrors cold — while not letting the coolers' vibrations creep in the opposite direction.

The prospects for a large, Japanese-led gravitational-wave detector suddenly improved towards the end of that decade, when Kajita stepped in to champion such a project. He had led a rebuilt Super-Kamiokande that made major breakthroughs in neutrino science, and he lent his credibility to KAGRA as someone who knew how to manage a big-science project. It was a similar role to the one that Barry Barish, a physicist at the California Institute of Technology (Caltech) in Pasadena, had for LIGO, says University of Tokyo physicist Shinji Miyoki.

In 2010, Japan's parliament approved funding for the project, which also has funding partners including South Korea and Taiwan. KAGRA's name — chosen from 600 suggestions made by the public — is also a reference to kagura, a dance for the gods that is part of Japan's ancient Shinto tradition, says Miyoki, who has worked on TAMA and CLIO, and now has a leading role at KAGRA.

A PROBLEM OF LOCATION

Construction began in 2012, and KAGRA's 6 kilometres of tunnels were dug in less than 2 years.

But the location also poses a problem: the mountain's rock is porous and soaked in water. KAGRA physicist Keiko Kokeyama recalls visiting the site in 2014, when she was working on LIGO at Caltech. "Inside the tunnel, it was raining very hard," she says, and the floor was covered in mud. Keeping the tunnels dry required an extra layer of lining, says Kokeyama, who oversees the interferometer's laser source, as well as having other roles.

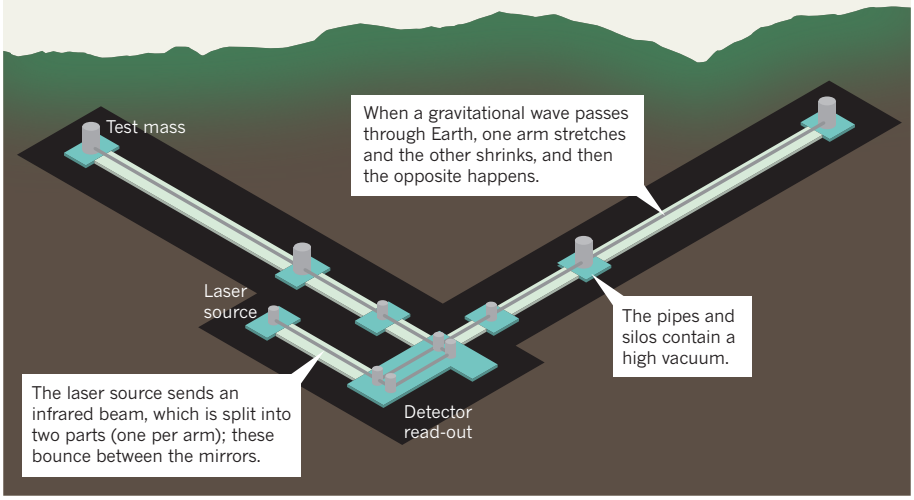
Each spring, when the snow above ground melts, the tunnels' drainage system takes away up to 1,000 tonnes of water per hour. This probably means that KAGRA will have to schedule yearly shutdowns in the wettest months, Kajita says. "With such conditions, I think it's unrealistic to operate."

The team hopes that the machine will be ready before the end of 2019, in time to join a year-long observation run that LIGO and Virgo are due to start in March. When KAGRA starts up, the world's gravitational-wave community will be watching. LIGO is planning an upgrade called LIGO Voyager that will also have cold mirrors — although not as cold as KAGRA's. And the US community is designing a 40-kilometre cryogenic machine called Cosmic Explorer. Meanwhile, researchers in Europe hope to build an observatory called the Einstein Telescope, which will be a 10-kilometre-sided triangle, and will be both cryogenic and underground. "I believe people will learn from KAGRA," says Kajita.

"It's a daunting but courageous thing that KAGRA has done," Shoemaker says, "to pursue several of the things we think we need for future detectors. It will help us enormously to have that experience." ■

JAPAN'S WAVE HUNTER

The Kamioka Gravitational Wave Detector (KAGRA) is the world's fourth major gravitational-wave detector — and Asia's first. Due to open in late 2019, it is the first one to be built underground, and to have cryogenically cooled mirrors, operating at around 20 kelvin. Both innovations should help KAGRA to separate the cosmic ripples from background noise.



© 2019 Springer Nature Limited. All rights reserved.

SOCIETY

Scientists under siege amid Nicaragua's political unrest

Government's response to the protests has led to the firing of university faculty members.

BY MICHELE CATANZARO



INTI OCON/AFP/GETTY

Students from across Nicaragua gather to demand the resignation of President Daniel Ortega.

Ongoing protests against the Nicaraguan government have led to violent clashes, and the crackdown by security forces has engulfed the country's scientists, causing some to flee their homes in fear for their lives.

The student-led protests started in April 2018 in response to a decree from President Daniel Ortega that increased social-security taxes and reduced pensions. Ortega's increasingly authoritarian administration tried to quell the protests with deadly force, which sparked demonstrations across Nicaragua. Fierce confrontations between protesters, police and activists supporting the government have resulted in more than 300 deaths.

Universities have fired faculty members who have criticized the administration's response to the demonstrations, and scientific conferences have been moved or postponed. In early December, the government shut down the offices of nine non-governmental organizations (NGOs), including the Rio Foundation in San Carlos, which focuses on environmental protections for the southeastern region of Nicaragua. The state seized the property of all nine NGOs, and they cannot operate legally in the country.

An economic crisis that has proliferated in the wake of the political unrest resulted in emergency cuts to the 2018 budget in August. They included a roughly 7% reduction for the National Council of Universities, Nicaragua's governing body for higher-education institutes.

The trouble has even affected the Nicaraguan Academy of Sciences, which released several statements starting in April in support of students and academic freedom. Its president, lawyer María Luisa Acosta, fled the country in May after receiving death threats. The threats stemmed from those statements of support and her long-standing criticism of government projects that would affect Indigenous groups and the environment.

CRACKDOWN

At least 40 faculty and staff members have been fired, and 82 students expelled, from the National Autonomous University of Nicaragua (UNAN), which has campuses in León and Managua, according to a report by Scholars at Risk, a New York City-based organization of universities and associations focused on protecting academic freedom.

They include Mauricio Álvarez Argüello, a biologist at UNAN in León, who was fired in November and allegedly attacked by police officers close to his house. Álvarez Argüello refused to sign a letter denouncing the actions of a government critic, and his brother is a constitutional scholar who has criticized Ortega's administration.

UNAN also fired Javier Pastora, former head of the surgery department and a surgeon at the university hospital in León. "I had worked at the hospital for 32 years," says Pastora.

A letter notifying Pastora of his dismissal gave no reason for the action. And UNAN officials did not respond to *Nature*'s request for comment. But Pastora attributes his sacking, as well as that of 12 other doctors and hospital staff who were fired at the same time, to the fact that they joined some of the protests. The physicians were also outspoken about alleged attempts by government forces to discourage wounded protesters from seeking treatment, by sending police to hospitals to arrest them.

The firings have also cut into collaborations with medical faculty in the United States. "It's so sad to see what has happened. I do not have a contact any more to work with," says Michael Lawson, a clinical researcher at the University of California, Davis. He has worked with Pastora since 2009 to set up an endoscopy department at the UNAN university hospital. Lawson provided donated equipment, and helped with training, surgery and an exchange programme for medical students.

"Students are a target for the government," says Jorge Huete-Pérez, a molecular biologist and senior vice-president of the University of Central America in Managua. Government security forces are now patrolling many university campuses, so a lot of students simply don't attend classes any more, he says.

CHANGE OF PLANS

The violence is also affecting scientific conferences. A November meeting of the Mesoamerican Society for Biology and Conservation, involving close to 1,000 scientists, had to be moved from Granada, Nicaragua, to Panama for security reasons. And the biannual Nicaraguan Biotechnology Conference, scheduled for September, has been postponed by a year.

Many people in Nicaragua fear that the situation will get worse. Several members, including Acosta, of the Civic Alliance — an organization that was meant to establish a dialogue with the government in the name of the protesters — have been arrested, threatened with death, or have fled the country.

Still, not everything is lost, says Huete-Pérez. "International participation can help a lot to create the environment for the government to sit down and negotiate." ■

© 2019 Springer Nature Limited. All rights reserved.



SPL

A lanthanum-based compound seems to act as a superconductor at near room temperature.

PHYSICS

Hint seen of new superconductor

Physicists might be a step closer to achieving their dream of superconductors that work at room temperatures.

BY DAVIDE CASTELVECCHI

Physicists say they might have achieved one of the most coveted goals of their discipline: the creation of a superconducting material that works at near room temperature.

The evidence is still preliminary and comes with a major caveat. So far, the material they created can exist only under pressures of about 200 gigapascals — or 2 million atmospheres.

But if confirmed, the feat would be the first example of superconductivity above 0 °C. Some physicists say that the work could be a milestone in the study of the property, which researchers hope will one day make the generation, transmission and use of electricity vastly more efficient.

'LONG-HELD DREAM'

"The observation is amazing," says Yanming Ma, a physicist at Jilin University in Changchun, China, although he cautions that the work is still in its early stages. Getting to room temperature has been "a long-held dream", Ma says, ever since superconductivity was discovered more than a century ago.

Russell Hemley, a materials chemist at the George Washington University in Washington DC, first announced evidence of superconductivity at −13 °C in May, and then revealed hints of an even higher 7 °C transition at a conference in August. His team is now publishing the results in *Physical Review Letters*[1].

The authors report seeing a sudden drop in electrical resistance at around 7 °C in a material they synthesized. The material is a 'superhydride' — a compound that contains a large amount of hydrogen — of lanthanum, with the chemical formula LaH_{10}.

The drop in resistance is the hallmark of a phase transition to superconductivity that occurs when the material is cooled below a threshold temperature. "We're very confident that we see a transition," says Hemley.

The achievement of superconductivity above 0 °C has no particular physical meaning, but it is "enormously important psychologically", says Mikhail Eremets, a physicist at the Max Planck Institute for Chemistry in Mainz, Germany. In 2014, Eremets' team showed that another hydrogen compound — hydrogen sulfide — becomes a superconductor at what was, at the time, the record high temperature of −83 °C (ref. 2).

RECORD HIGHS

In their experiment, Hemley and his collaborators placed a diamond anvil in a synchrotron beamline at the Argonne National Laboratory near Chicago in Illinois. They used the anvil's diamond tips to squeeze a minuscule sample of lanthanum and hydrogen to pressures of up to 200 gigapascals.

Next, they temporarily heated the compound and watched its structure change along with its conductive properties, monitoring the process using X-ray diffraction. The researchers managed to produce a new crystal structure — LaH_{10} — which previous simulations by their team and others, including Ma's, suggested would be superconducting at very high temperatures.

They then allowed it to cool while keeping it at high pressure, and measured its electronic properties. In certain conditions, they saw the electrical resistance drop at a temperature of 280 kelvin, or about 7 °C.

The evidence presented in Hemley's paper has yet to convince Eremets. Follow-up experiments in his own lab suggest that the material's transition temperature is not quite as high as that, although it still comes in at an impressive −23 °C (ref. 3).

But Hemley says that in as-yet-unpublished follow-up work, his team detected another important sign of superconductivity: the material expelled existing magnetic fields from itself. The phenomenon is considered to be gold-standard evidence of superconductivity and, if confirmed, could clinch the team's claim.

JUST THE BEGINNING

Compounds such as the lanthanum superhydride made by Hemley's team, and the hydrogen sulfide studied by Eremets in 2015, are conventional superconductors, meaning that their physical properties have been well understood since the 1950s.

Conventional superconductors with room-temperature transitions have been predicted for several decades[4], but only recently have these predictions begun to be tested in the lab.

Hemley is confident that other materials exist — beyond even those explored in simulations — with even higher transition temperatures.

More-exotic superconductors discovered since the 1980s had, until Eremets' 2014 results, boasted record-high transition temperatures, but are yet to achieve room-temperature superconductivity. Their theoretical underpinnings are still unexplained.

Hemley says that his team's experiments could offer hints on how to develop materials that might have similar electronic properties at less extreme pressures. "This is just the beginning of a new era of superconductivity," says Hemley. ■ SEE NEWS FEATURE, P.15

1. Somayazulu, M. *et al. Phys. Rev. Lett.* (in the press).
2. Drozdov, A.P., Eremets, M. I. & Troyan, I. A. Preprint at https://arxiv.org/abs/1412.0460 (2014).
3. Drozdov, A. P. *et al*. Preprint at https://arxiv.org/abs/1812.01561 (2018).
4. Ashcroft, N. *Phys. Rev. Lett.* **92**, 187002 (2004).

CORRECTION

The News Feature 'Africa's silent epidemic' (*Nature* **564**, 24–26; 2018) erroneously referred to HIV as a DNA virus. It is, in fact, a retrovirus.

© 2019 Springer Nature Limited. All rights reserved.



DESIGN PICS/NGC

RESEARCH

What to watch for in 2019

Climate research, open access and a biosafety rethink are set to shape the year's science.

Elephant seals carrying sensors will help researchers to gather ocean data as part of a massive mission to study Antarctica's Thwaites glacier.

POLAR PROJECTS

In January, US and UK researchers will descend on Antarctica to begin their largest joint mission to the continent in more than 70 years. The aim of the five-year project is to understand whether the remote and seemingly unstable Thwaites Glacier will start to collapse in the next few decades. It includes efforts to study ocean conditions near the Florida-sized glacier using autonomous underwater vehicles and sensors affixed to seals. Later in the year, European scientists plan to start drilling into the ice sheet on Antarctica's Little Dome C in a quest to recover a 1.5-million-year-old ice core. If they're successful, the core will yield the oldest pristine record of climate and atmospheric conditions.

BIG BUCKS

China could emerge as the world's biggest spender on research and development, after adjusting for the purchasing power of its currency, once countries publish their 2018 spending data in late 2019. Outlays on science in China have accelerated since 2003, although the country still trails behind the United States on measures of research quality. Over in Europe, officials will try to agree on how to disburse a proposed €100 billion (US$110 billion) through the European Union's next research-funding programme, Horizon Europe, which begins in 2021. It's unclear how fully UK researchers will be able to participate, as uncertainty over Brexit continues to plague the country.

HUMAN ORIGINS

More fossils illuminating the origins of ancient hominin species could emerge from islands in southeast Asia — a region of intense interest since archaeologists discovered a human-like 'hobbit' species on the Indonesian island of Flores in 2003. Ongoing digs could reveal more about the first human inhabitants of the Philippine island of Luzon, including whether their isolation led to a diminutive stature, similar to what seems to have occurred on Flores.

COLLIDER CRUNCH

It could be a make-or-break year for plans to build a successor to the Large Hadron Collider (LHC). Physicists in Japan proposed hosting the roughly US$7-billion International Linear Collider (ILC) in 2012, after scientists at the LHC in Geneva, Switzerland, announced the discovery of the Higgs boson. The ILC would study the Higgs in detail. But a 2018 report commissioned by the Japanese government failed to support the project, citing its cost. Japan is the only country that has shown interest in hosting the ILC, and the government is expected to issue a statement on whether it will do so by 7 March.

GENE-EDITING FALLOUT

Geneticists will continue to deal with the repercussions of last year's claim by He Jiankui to have helped produce the world's first gene-edited babies. Researchers hope to confirm whether He, a genome-editing researcher at the Southern University of Science and Technology in Shenzhen, China, modified the genes of two embryos that produced twin girls. Following an international outcry, scientists will attempt to uncover any potential side effects of the process, and create a framework to ensure that any future efforts to edit heritable human DNA — such as that in eggs, sperm or embryos — happen in a responsible and regulated way.

PLANNING FOR PLAN S

Subscription journals could shift their business models to accommodate Plan S, the effort to flip scholarly publications to a fully open-access model. Publishers have a year before the scheme's backers will require the researchers they fund to immediately archive papers accepted for publication in free-to-access repositories — a practice that many journals currently forbid. The drive for open science also underpins a 2019 effort by funders and research organizations in the Netherlands that seeks to move away from using citations and impact factors to assess researchers. ▶

© 2019 Springer Nature Limited. All rights reserved.

BIOSAFETY BIBLE

The World Health Organization expects to finish a major revision of its *Laboratory Biosafety Manual* in mid-2019. The widely used guidelines outline best practices for the safe handling of pathogens such as Ebola. This is the manual's first overhaul since 2004. The revisions will increase the focus on creating site- and experiment-specific risk assessments, and on improving the management, practices and training of lab personnel. The rethink aims to discourage labs from approaching biosafety by rote, and encourage the creation of more flexible and effective procedures.



CARLOS OSORIO/REUTERS

Canada is set to harvest the results of a boom in marijuana research.

CLIMATE TINKERING

As carbon emissions continue to rise, 2019 could see the first experiments that are explicitly aimed at understanding how to artificially cool the planet using a practice called solar geoengineering. Scientists behind the Stratospheric Controlled Perturbation Experiment (SCoPEx) hope to spray 100-gram plumes of chalk-like particles into the stratosphere to observe how they disperse. Such particles could eventually cool the planet by reflecting some of the Sun's rays back into space. Geoengineering sceptics worry that the practice could have unintended consequences and distract from efforts to reduce greenhouse-gas emissions. The US-led SCoPEx team is awaiting the go-ahead from an independent advisory committee.

HIGH HOPES

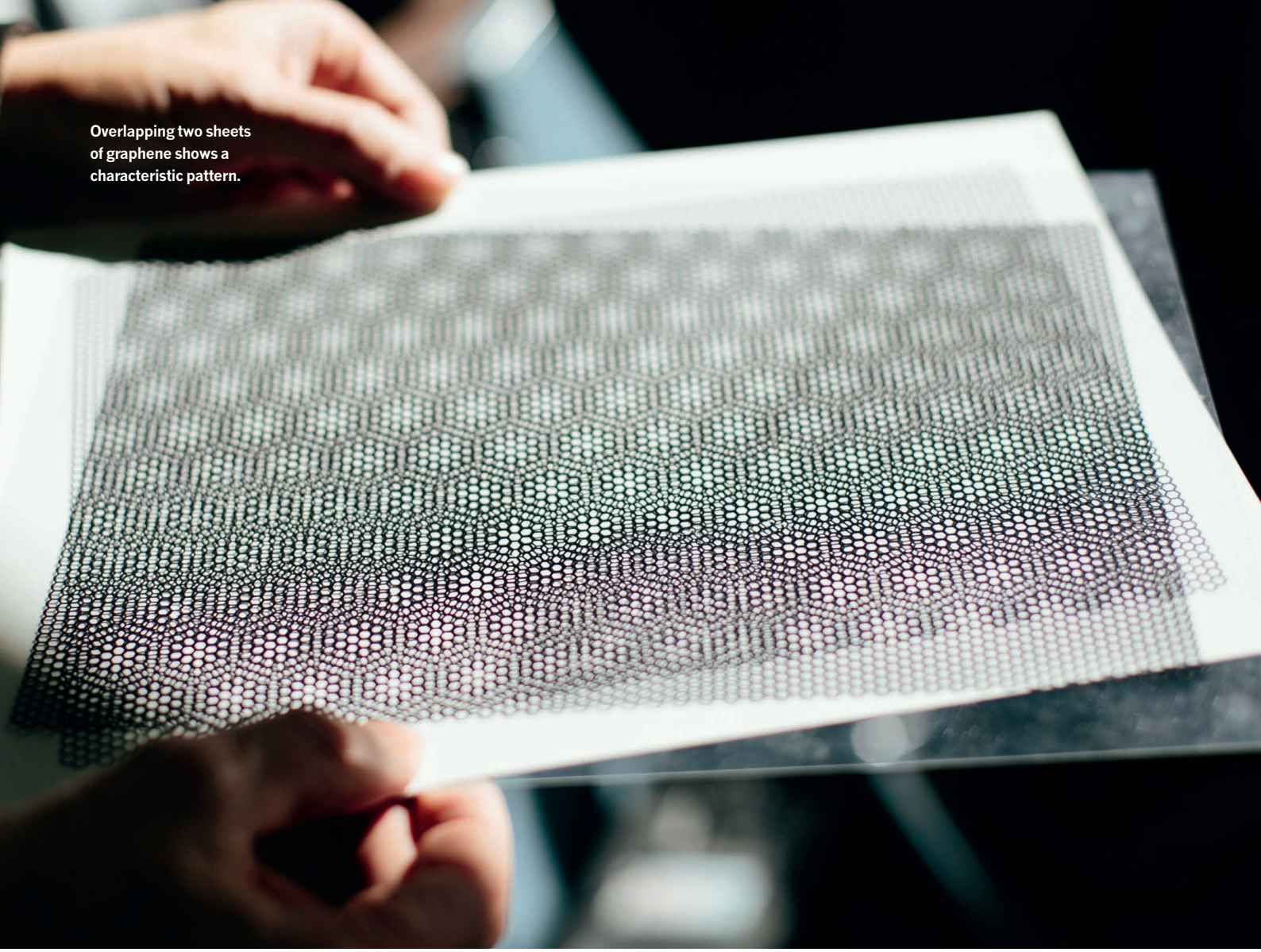
Researchers in Canada should start to see the first results from a flurry of studies into the cultivation and basic biology of cannabis. In October, the country legalized the plant for all uses — the second nation in the world, after Uruguay, to do so — leading to funding windfalls for marijuana research from provincial and federal governments. By the end of 2019, researchers at the University of Guelph hope to launch Canada's first dedicated academic centre for cannabis research, which will study everything from the plant's genetics to its health benefits.

COSMIC SIGNALS

The world's largest radio telescope — China's Five-hundred-meter Aperture Spherical Radio Telescope — should be fully operational and available to researchers from September. Since the start of its commissioning phase in 2016, the 1.2-billion-yuan (US$170-million) mega-telescope has spotted more than 50 new pulsars: dense, rapidly spinning dead stars. It will soon hunt for the faint signals that emerge from phenomena such as fast radio bursts and clouds of cosmic gas. Meanwhile, astronomers will decide whether to press ahead with building the Thirty Meter Telescope on the Hawaiian mountain Mauna Kea. In 2018, the plans cleared the last of a long series of legal challenges lodged by locals. ■

COMPILED BY ELIZABETH GIBNEY

© 2019 Springer Nature Limited. All rights reserved.



Overlapping two sheets of graphene shows a characteristic pattern.

JULIETTE HALSEY FOR *NATURE*

SUPERCONDUCTIVITY WITH A TWIST

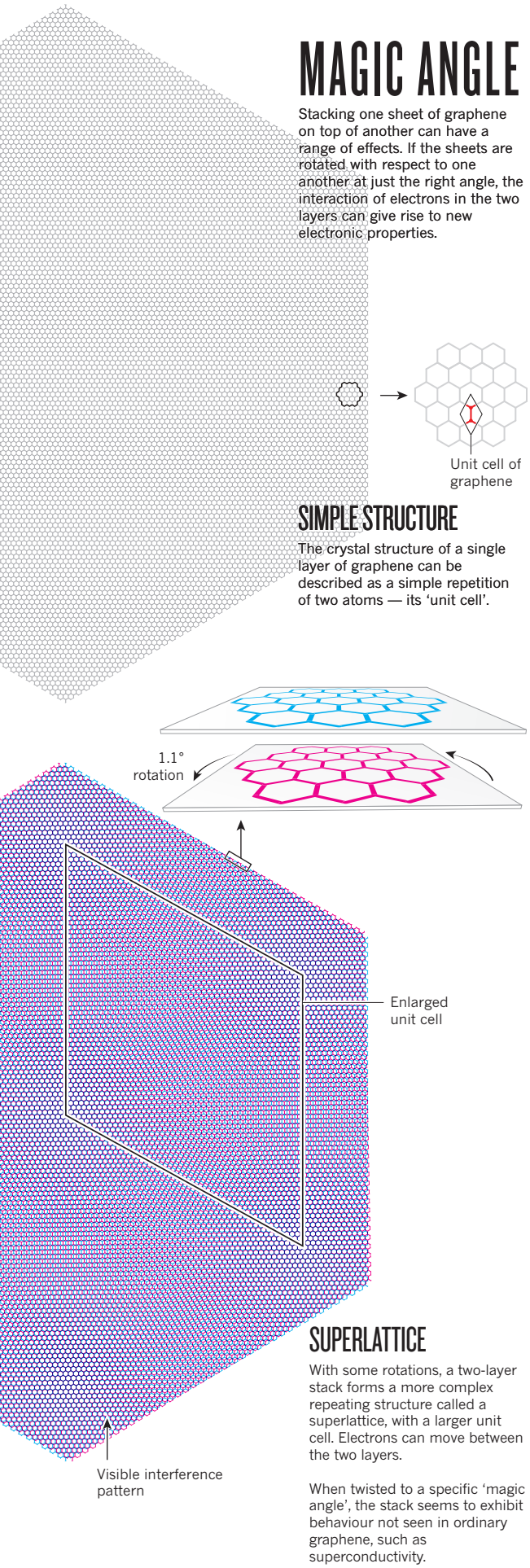
RESEARCHERS ARE SCRAMBLING TO UNDERSTAND CURIOUS BEHAVIOUR IN MISALIGNED STACKS OF GRAPHENE.

BY ELIZABETH GIBNEY

It was the closest that physicist Pablo Jarillo-Herrero had ever come to being a rock star. When he stood up in March to give a talk in Los Angeles, California, he saw scientists packed into every nook of the meeting room. The organizers of the American Physical Society conference had to stream the session to a huge adjacent space, where a standing-room-only crowd had gathered. "I knew we had something very important," he says, "but that was pretty crazy."

The throngs of physicists had come to hear how Jarillo-Herrero's team at the Massachusetts Institute of Technology (MIT) in Cambridge had unearthed exotic behaviour in single-atom-thick layers of carbon, known as graphene. Researchers already knew that this wonder material can conduct electricity at ultra-high speed. But the MIT team had taken a giant leap by turning graphene into a superconductor: a material that allows electricity to flow without resistance. They achieved that feat by placing one sheet of graphene over another, rotating the other sheet to a special orientation, or 'magic angle', and cooling the ensemble to a fraction of a degree above absolute zero. That twist radically changed the bilayer's properties — turning

© 2019 Springer Nature Limited. All rights reserved.



it first into an insulator and then, with the application of a stronger electric field, into a superconductor.

Graphene had previously been cajoled into this behaviour by combining it with materials that were already known to be superconductors, or by chemically splicing it with other elements. This newfound ability to induce the same properties at the flick of a switch turned heads. "Now you put two, non-superconducting atomic layers together in a certain way and superconductivity pops up? I think that took everyone by surprise," says ChunNing Jeanie Lau, a physicist at the Ohio State University in Columbus.

Physicists at the meeting were even more excited because of the way in which a graphene bilayer seems to become a superconductor. There were hints that its remarkable properties arose from strong interactions or 'correlations' between electrons — behaviour that is thought to underlie bizarre states of matter in more-complex materials. Some of those materials, namely ones that superconduct at relatively high temperatures (although still well below 0 °C), have baffled physicists for more than 30 years. If superconductivity in simple graphene is caused by the same mechanism, the material could be the Rosetta stone for understanding the phenomenon. That, in turn, could help researchers to engineer materials that superconduct close to room temperature, which would revolutionize many areas of modern technology, including transportation and computing.

"Immediately I could see pretty much everyone I know become really excited," says Lau. But while she listened in amazement to the talk, others couldn't wait. Andrea Young, a condensed-matter physicist at the University of California, Santa Barbara, had left the meeting to rush back to his laboratory. His team was one of a handful around the world already exploring twisted graphene, looking for hints of recently predicted strange behaviour. Young scanned the *Nature* papers[1,2] from the MIT group, which were published two days ahead of the talk, and found what he needed to know to replicate the experiment. That turned out to be harder than anticipated. But by August, having joined forces with a group at Columbia University in New York City led by physicist and friend Cory Dean, he and his team succeeded[3]. "We had reproduced it many times ourselves," says Jarillo-Herrero. But having the confirmation of a second group, he says, "was tremendously reassuring".

Although the Young and Dean collaboration was the first to publicize its replication results, activity behind the scenes is frenetic, says Lau. "I haven't seen this much excitement in the graphene field since its initial discovery," she says. Three other teams told *Nature* that they have replicated some or all of the MIT findings, although some are keeping their cards close to their chests while they experiment with other 2D materials and tweak layers in new ways, looking for other displays of strong electron interactions. "Everyone is taking their favourite thing and twisting it with their other favourite thing," says Young. Meanwhile, theorists trying to explain the behaviour have posted more than 100 papers on the topic to the arXiv preprint server. But sorting out whether the same mechanism that underlies superconductivity in high-temperature superconductors is at play in twisted graphene will take much more information, says Lau. "So far, apart from the fact that this is a really interesting system," she says, "I don't think the theorists agree on anything."

FINDING THE MAGIC

The audience at Jarillo-Herrero's talk in Los Angeles was excited but also sceptical. Conference delegates teased him that the last time someone had presented something so cool, it was Jan Hendrik Schön, whose string of dazzling results on superconductivity and other phenomena turned out to be fraudulent. "They were joking," Jarillo-Herrero says, "but they said they'd need to see this reproduced before they would believe it."

Although twisted graphene's superconducting behaviour came as a surprise, the idea that something intriguing could happen was not. Overlaid at angles of more than a few degrees, two graphene sheets usually behave independently. But at smaller angles, the misalignment of the two lattices can create a 'superlattice' in which electrons can move between layers. Theorists had predicted[4,5] that at specific small twists — magic angles — the underlying structure of the superlattice would drastically change the behaviour of electrons, slowing them down and enabling them to interact in ways that change the material's electronic properties (see 'Magic angle'). In theory, all kinds of layered 2D material, when twisted to the proper angle, can form such superlattices. But no one knew how a material's properties might change, or at what angle such a change might occur.

Back in 2010, Eva Andrei, a physicist at Rutgers University — New Brunswick in New Jersey, and her colleagues saw hints of strange behaviour in graphene[6] around the same magic angle later observed by Jarillo-Herrero and his team, but many doubted whether the theory worked at all. "I didn't believe it, says Philip Kim, an experimental

© 2019 Springer Nature Limited. All rights reserved.



JULIETTE HALSEY FOR *NATURE*

Physicist Pablo Jarillo-Herrero (far left) with three graduate students in his lab at the Massachusetts Institute of Technology.

physicist at Harvard University in Cambridge, Massachusetts. "But I admit I was completely wrong," he says.

When Young arrived back at his lab in March, he thought that reproducing the MIT group's results seemed trivial, he says. Young's team could achieve the very low temperatures needed, and the researchers were already experts in preparing very clean samples. But coaxing graphene sheets to align at just the right angle — a twist of around 1.1° — turned out to be a struggle.

Hitting the angle is difficult, not least because it subtly changes from sample to sample, depending on how each one is made. "You have to do some searching," says Andrei. Moreover, because twisted graphene's structure is so close to that of graphite, in which successive layers are all oriented in the same direction, the slightest heat or strain can cause the layers to fall into alignment. "It doesn't want to stay where you put it," says Young.

Dean's lab, which was also working on the problem, hit on a solution: when the team overshot the twist in a number of devices, at least some samples would settle at the magic angle as they rotated back towards alignment. But getting those samples to superconduct required equipment that could reach a fraction of a degree above absolute zero, which his lab lacked. Working with Young's team, the researchers soon measured several devices in which resistance shot up — characteristic of an insulator — but dropped to zero, as in superconductors, when they fed in more electrons by applying an electric field.

It is the only other team apart from Jarillo-Herrero's to publish its findings so far, but that will not be the case for long, says Andrei. "Everyone I know is working on this," she says.

SOMETHING UNCONVENTIONAL

One reason for the intense interest in twisted graphene is the stark similarities between its behaviour and that of unconventional superconductors. In many of these, electric current runs without resistance at temperatures well above what the conventional theory of superconductivity generally allows. But quite how that happens remains a mystery: one that, when solved, could allow physicists to engineer materials that conduct electricity with zero resistance near room temperature. Achieving that could enable radically more-efficient transmission of electricity, and, by slashing energy costs, allow superconductors to find uses in a host of new technologies.

All forms of superconductivity rely on electrons pairing up in ways that allow them to travel without resistance. In conventional superconductors — the kind that power the magnets in magnetic resonance imaging (MRI) machines — electrons pair up only indirectly, as a by-product of the interplay between the particles and vibrations in their atomic lattice. Electrons ignore their fellows, but end up thrown together in a way that helps them to navigate without resistance at temperatures a few degrees above absolute zero. But in unconventional superconductors — many of which carry current with zero resistance at closer to 140 kelvin — electrons seem to pair up through a direct and much stronger interaction.

The MIT experiments showed hints of this unconventional superconductivity. Although twisted bilayer graphene became superconducting only at extremely low temperatures, it did so with very few freely moving electrons. That suggests that, unlike in a conventional superconductor, whatever force drew the electrons together must be relatively strong. The proximity of the superconducting state to an insulating one also mirrors what is seen in a group of high-temperature superconductors made from ceramics, called cuprates. In those systems, the zero-resistance state often borders a 'Mott' insulator — in which no current flows, despite the presence of free electrons, because mutual repulsion between the particles pins them in place.

If the same mechanisms are at play in twisted bilayer graphene, it could be boon to theorists. One problem with cuprates, such as yttrium barium copper oxide, is that they are a jumble of elements that proves difficult to model. "The hope is of finding the same phenomenology but in a much simpler system, one which theorists can stick their teeth into and make some progress," says Andrei.

Graphene is also an experimentalist's dream. Studying the switch to superconductivity means measuring what happens as more electrons are added to the material. In cuprates, this is done by inserting atoms of a different element into the material — a process known as doping — which means making an entirely new sample for each point on a chart. In twisted graphene, however, researchers can make the switch simply by turning a knob on a voltage source, says Andrei. "This is a huge benefit."

No one knows yet whether twisted graphene is really acting like an unconventional superconductor, or even whether the behaviour arises exactly because of the conditions described by the magic-angle theory. The flood of theory papers posted since March covers every possibility. Because correlated systems such as those seen in twisted graphene are too complex to calculate in their entirety, theorists use approximations that differ from model to model. That makes theories flexible enough for physicists to sometimes tweak them to fit new data, says Young. Few theories explain the findings in full, and many do not include predictions that would allow experimentalists to tease apart different scenarios, adds Jarillo-Herrero. For "an experimentalist like me they all seem similarly sensible", he says. "I'm a bit disoriented in theory land."

© 2019 Springer Nature Limited. All rights reserved.

So far, there is evidence for both unconventional and conventional superconductivity in graphene. As-yet-unpublished data from the MIT group suggest that other phenomena seen in unconventional superconductors are also present in the material, says Jarillo-Herrero. For one thing, his team has observed that the strength of the magnetic field necessary to strip superconductivity from a sample, through a process known as the Meissner effect, varies with direction (it should be uniform in conventional superconductors).

CAUTIOUS APPROACH

But results from Young and Dean's groups suggest more caution is needed. Their samples are more uniform than those of the MIT group, says Young, and show some contrasting results. In particular, superconductivity appears when the number of electrons is turned down but not when it is turned up, an asymmetry that is arguably more consistent with a conventional superconductor. And, in contrast to cuprates, which can be insulating at higher temperatures than those at which they superconduct, in twisted graphene the two states seem to be present in a similar temperature range, he adds. Further tests, such as seeing whether the superconducting state still occurs when experimentalists restrict vibrations in the sample but still allow electron interactions, could help to clarify the situation, says Young. Andrei's group is also working on imaging the material at the atomic level, to reveal effects that could be washed out when studying the sample as a whole. Andrei says her team's preliminary data have revealed new phenomena that could help to make sense of the underlying physics, although she is so far unwilling to give any more away.

"EVERYONE IS TAKING THEIR FAVOURITE THING AND TWISTING IT WITH THEIR OTHER FAVOURITE THING."

Understanding the outcomes of experiments — along with devising set-ups that work well on 2D materials — can be a challenge. In this delicate system, Young says that even the material used to make the electrodes can interfere with results. "You have to be careful about interpreting what you see, because you don't know what's an intrinsic property of the system and what's an effect of the set-up." Young says the mechanism behind the superconductivity could well turn out to be conventional, but that it is exciting even if it doesn't help to explain high-temperature superconductivity. "This is already one of the coolest results to come out of this field in the past ten years," he says.

Regardless of whether it resembles exotic forms of superconductivity, researchers say the system is fascinating because it is a rare example of dramatic change coming from a small physical tweak. "That fact alone is pretty amazing and remarkable," says Dean. "What is it about this system that gives rise to superconductivity that is absent away from this precise twist angle?"

Whatever is going on in the superconducting state, physicists agree that the accompanying insulating state is almost impossible to explain without some kind of interaction between electrons. Like a metal, graphene is ordinarily conductive, with free electrons that interact only with the atomic lattice and not with one another. Somehow, despite the presence of these free electrons, which are absent in conventional insulators, bilayer graphene can block the flow of electricity, suggesting that interactions are at play.

This is exciting because electron interactions underlie many of the weird and wonderful states of matter that have been uncovered over the past few decades. These include quantum spin liquids — strange disordered states in which electrons' magnetic fields never align — and fractional quantum Hall states, phases of matter defined by topology, a previously unknown kind of unifying property that might be harnessed to build extremely robust quantum computers. "Understanding strongly correlated systems is where a lot of the big questions, and also perhaps big opportunities, are in condensed-matter physics right now," says Young. Many of these states emerge under conditions that, at least to electrons, look similar to those that arise in graphene at the magic angle. This raises the possibility that other intriguing states could emerge from twisted bilayers, says Rebeca Ribeiro-Palau, a physicist at the Centre for Nanoscience and Nanotechnology in Palaiseau, France, and formerly a postdoc in Dean's lab. "For me, the presence of a superconducting state is a symptom of something more interesting," she says.

Crucially, graphene and other 2D systems allow for much greater experimental control than do other strongly correlated materials, she says. Researchers can smoothly tune not only the electric field to alter behaviour, but also the twist angle — while at Columbia, Ribeiro-Palau and her colleagues used the tip of an atomic force microscope to smoothly spin one layer with respect to the other[7]. As has been demonstrated by the Young and Dean collaboration, experimentalists can also fine-tune the distance between layers by applying pressure. Squeezing the layers closer together increases the strength of the interaction between electrons in the sheets, a boost that means magic-angle conditions can happen at much bigger — and more stable — rotations.

DOING THE TWIST

Kim and his colleagues have already replicated the graphene finding, he says. Now they are looking to see whether they can also generate superconductivity or perhaps magnetism in twisted layers of more-complex 2D semiconductors, called transition-metal dichalcogenides. Before the MIT result, Kim's was one of a few teams that was already probing the effects of rotating one 2D layer on top of another, a nascent area of research sometimes known as twistronics. With the possibilities demonstrated in graphene, the idea is now taking off. "In principle, you can apply the concept to all the 2D materials and twist to see what happens," says Kim. "There is the possibility that you find something completely unexpected."

Meanwhile, Feng Wang at the University of California, Berkeley, says he and his colleagues have seen signs of superconductivity in triple-stacked layers of graphene even without a twist. Layering three sheets in a particular orientation[8] achieves a superlattice geometry similar to that in magic-angle twisted bilayers, and results in similarly strongly correlated physics, he says.

Physicists are optimistic that the crossover between two previously separate fields — 2D materials and strongly correlated systems — will lead to exciting results. "It's giving us an opportunity to talk to a whole community of people we haven't had the chance to talk to in the past," says Dean. And applied physicists are thinking about how the unusual properties of twisted 2D stacks might be harnessed to store and process information in super-efficient ways. Rotating or squeezing materials could also become a new way to switch an electronic device's behaviour.

But for now, many researchers are focused on sorting out the fundamentals. This month, experimentalists and theorists will gather at the Kavli Institute for Theoretical Physics in Santa Barbara for a workshop that will thrash out key questions in the burgeoning field. Jarillo-Herrero hopes the meeting will help bring theorists into alignment. "At the moment, they can't even agree on the basics." By then, more experimentalists might be willing to show their hand and publicly reveal their data, he adds.

Even though physicists don't know how significant the discovery will ultimately be, Young says there's a takeaway message from the dozens of theory papers that have appeared since the MIT publications: "Anything could come out of this, and something certainly will." ■

Elizabeth Gibney *is a senior reporter for* Nature *in London.*

1. Cao, Y. *et al. Nature* **556**, 43–50 (2018).
2. Cao, Y. *et al. Nature* **556**, 80–84 (2018).
3. Yankowitz, M. *et al.* Preprint at https://arxiv.org/abs/1808.07865 (2018).
4. Bistritzer, R. & MacDonald, A. H. *Proc. Natl Acad. Sci. USA* **108**, 12233–12237 (2011).
5. Suárez Morell, E., Correa, J. D., Vargas, P., Pacheco, M. & Barticevic, Z. *Phys. Rev. B* **82**, 121407 (2010).
6. Li, G. *et al. Nature Phys.* **6**, 109–113 (2010).
7. Ribeiro-Palau, R. *et al. Science* **361**, 690–693 (2018).
8. Chittari, B. L., Chen, G., Zhang, Y., Wang, F. & Jung, J. Preprint at https://arxiv.org/abs/1806.00462 (2018).

© 2019 Springer Nature Limited. All rights reserved.

COMMENT

EXHIBITIONS Must-see shows of 2019 include Leonardo a gogo **p.22**



MUSEUMS Marking the anniversary of Alexander von Humboldt's birth **p.22**

NIH Survey reveals uneven enforcement of mandate on animal gender **p.25**

EQUITY Shouting match over gene drives drowns out key voices **p.25**

JONAS GRATZER/LIGHTROCKET/GETTY



The Tsho Rolpa valley in Nepal, where increased meltwater from glaciers in the Himalayas puts local communities at risk.

Collapsing glaciers threaten Asia's water supplies

Tracking moisture, snow and meltwater across the 'third pole' will help communities to plan for climate change, argue **Jing Gao** and colleagues.

The 'third pole' is the planet's largest reservoir of ice and snow after the Arctic and Antarctic. It encompasses the Himalaya–Hindu Kush mountain ranges and the Tibetan Plateau. The region hosts the world's 14 highest mountains and about 100,000 square kilometres of glaciers (an area the size of Iceland). Meltwater feeds ten great rivers, including the Indus, Brahmaputra, Ganges, Yellow and Yangtze, on which almost one-fifth of the world's population depends[1].

Climate change threatens this vast frozen reservoir (see 'Third pole warming'). For the past 50 years, glaciers in the Himalayas and Tibetan Plateau have been shrinking[2]. Those in the Tian Shan mountains to the north have lost one-quarter of their mass, and might lose as much as half by mid-century[3]. Their meltwater is expanding lakes[4]. River flows at the start of summer peak earlier than they did 30 years ago[5]. And weather patterns are shifting. A weaker Indian monsoon is reducing precipitation in the Himalayas[6] and southern Tibetan Plateau; snow and rain are increasing in the northwestern Tibetan Plateau and Pamir Mountains[2].

Researchers still don't understand why these changes vary so much across the region, or how they will pan out. Some rivers in central Asia, such as those feeding ▶

© 2019 Springer Nature Limited. All rights reserved.

the Aral Sea, are projected to gradually dry up[7]. Others — such as the upper Ganges, Brahmaputra, Salween and Mekong — are likely to swell, at least until 2050 (ref. 7).

Already, Tibetan communities are dealing with the impacts of collapsing glaciers. In October 2018, debris dammed the Yarlung Tsangpo River, which forms the headwater of the Brahmaputra, threatening areas as far afield as Bangladesh with flooding.

Communities need information to help them manage risks and water supplies. They need to know which glaciers are melting fastest, and how changing snowfall and a warmer climate are affecting the accumulation and disappearance of ice and the volumes of rivers and lakes.

MONITORING NETWORK

The water cycle is hard to monitor in this vast, high and remote region. Satellite images and climate models are too coarse to resolve local changes.

A network of monitoring stations is needed across the region. It must track classical meteorological variables, such as air temperature, humidity, air pressure, precipitation and winds. And it needs to expand data on the water cycle, by measuring the stable isotopes of hydrogen and oxygen in water vapour. This provides crucial insights into the origin of atmospheric moisture and the processes it has gone through, such as evaporation and condensation.

As a first step, an international scientific programme called the Third Pole Environment (TPE; led by T.Y.) has set up 11 ground stations and tethered balloons since 2014, working with the Institute of Tibetan Plateau Research, Chinese Academy of Sciences, in Beijing. This monitoring network is already larger than similar efforts in Antarctica and the Arctic, and almost doubles the number of such stations around the world.

But there is more to be done. Researchers need a better understanding of the relationships between the third pole's complex terrain and the weather patterns and processes that affect precipitation and ice-melting. The water cycle must be traced in three dimensions — as liquid water, ice and water vapour, on the ground and in the air — and changes monitored. Computer models also need to be tailored to provide accurate projections for the region.

NEED TO KNOW

Two weather patterns — the Indian monsoon and prevailing westerly winds — drive most of the moisture flow towards the third pole. As the Indian continent heats up in spring and summer, convection draws moisture northwards from the Bay of Bengal, Arabian Sea and Indian Ocean; this falls as precipitation in the Himalayas and beyond[8]. In the north and west of the region, strong westerly winds bring moisture from the Mediterranean. Across the whole region, water also evaporates from the soil and is given off by plants through transpiration.

"Satellite images and climate models are too coarse to resolve local changes."

We know these patterns thanks to observations of stable isotopes in water. In the vertical dimension, such data are revealing how moisture mixes in air masses and through processes at the atmospheric boundary layer. That information also records how moisture is released daily from glaciers, as their surfaces and the air heat and cool.

We still lack a quantitative understanding of the role of each process in the overall water budget. Nor is it clear how much water passes between solid, liquid and vapour phases, affecting regional hydrology. Physical processes that affect glaciers are poorly understood, including the impacts of aerosols and surface debris on ice accumulation and melting. We cannot predict how much meltwater will descend into lakes and rivers, nor how wet soils might increase local precipitation. The region's complex and varied topography is another confounding factor.

Then there is climate change. The westerly jet stream in East Asia has strengthened in winter for the past few decades[9], yet the Indian summer monsoon is weakening. Both trends affect the distribution of snowfall and thus the reflectivity (or albedo), energy budget (the balance of all the energy that enters and leaves the Earth system) and the water budget of the land's surfaces.

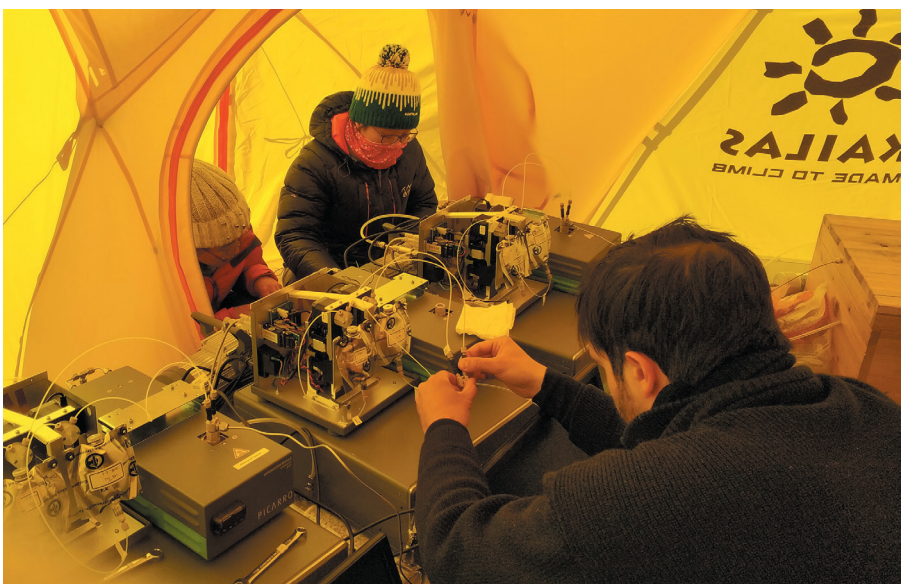
Global climate models are designed to simulate large-scale features of atmospheric circulation, and so struggle to reproduce weather patterns in the third pole. New models and data will be needed to improve them. Only 0.1% of glaciers and lakes in the region have monitoring stations. Few areas higher than 5,000 metres above sea level have weather stations, let alone water-isotope detectors.

NEXT STEPS

The first priority must be to extend the network of weather and isotope-monitoring stations. There are already plans to install 20 additional stations across a wider area of the third pole later this year; others can be added as learning progresses. The installation is part of China's Pan-TPE research programme, involving scientists from Norway to Nepal. It has a budget of 1.48 billion yuan (US$215 million) for 5 years to study environmental changes in the third pole, Iranian Plateau, Caucasus Mountains and Carpathian Mountains. Another programme — the Second Tibetan Plateau Scientific Expedition and Research (STEP) project — will receive 4.35 billion yuan over 5 years from 2019 to study environmental change in the Tibetan Plateau. Throughout this 10-year programme, the cost of instruments, staff and maintenance is likely to rise from 8 million yuan a year to about 150 million yuan.

Most of the monitoring stations will be set up along two axes. A south–north line of 15 stations at 100–500-kilometre intervals, stretching from the tropical Indian Ocean, across Bangladesh and Nepal to the Tian Shan Mountains, will monitor processes linked to the monsoon. An east–west transect, stretching from the Iranian Plateau to China's Loess Plateau, and comprising 12 stations 200–500 kilometres apart, will examine the impacts of the westerlies. Snow and rainfall, glacier melt and lake and river discharge will be observed in river basins, along with levels of glacial debris, permafrost and groundwater.

The interplay between elevation, atmospheric circulation and water vapour will be



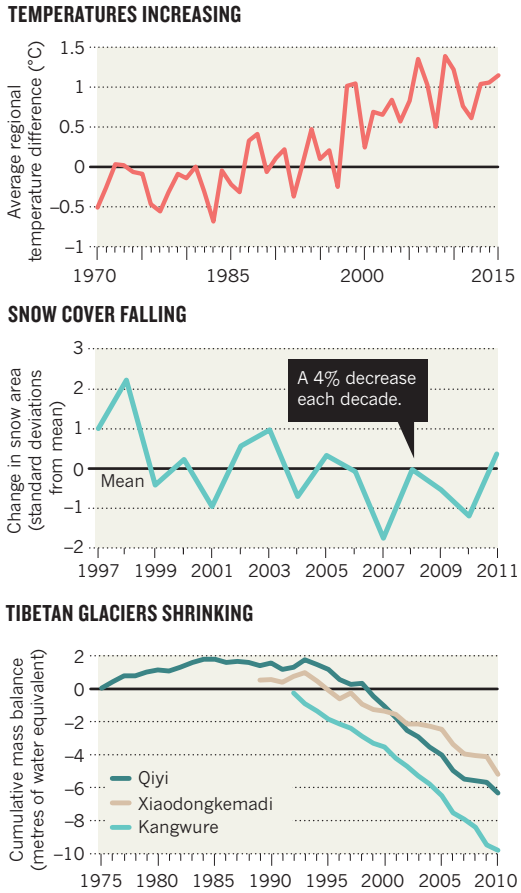
FEI LI

Researchers set up instruments to measure stable isotopes in atmospheric water vapour near Everest.

© 2019 Springer Nature Limited. All rights reserved.

THIRD POLE WARMING

Climate change is altering precipitation across the Himalayan mountain ranges and the Tibetan Plateau.





SOURCES: TEMPERATURES: T. YAO *ET AL. BULL. AM. METEOROL. SOC.* (IN THE PRESS); SNOW COVER: S. S. P. SHEN *ET AL. THEOR. APPL. CLIMATOL.* **120**, 445–453 (2015); GLACIERS: REF. 2; MAP: J. GAO, T. YAO & W. WANG.

tracked through hourly measurements at three hotspots: the Pamir Mountains (dominated by westerlies), the Himalayas (affected by the Indian monsoon) and the Hengduan Mountains (where the East Asian monsoon prevails). Each site will host 10 stations at 200-metre intervals in altitude.

Installing and maintaining this network will be challenging. Devices must be robust and equipped with the latest technologies, such as fast, laser-based spectroscopic isotope measurements and high-resolution light detection and ranging (lidar) systems. Instruments must be regularly calibrated. More than 200 professional staff members must be trained to run them.

The second priority is for data to be shared and fed into global and regional climate models. A new generation of Earth-system models should be developed for the third pole, representing its atmosphere, cryosphere, hydrosphere and biosphere. Models should include interactions at very high resolution and encompass water's stable isotopes, as well as aerosols and biogeochemical cycles.

These models should explore the regional implications of different scenarios of human activities and climate-mitigation strategies (greenhouse-gas emissions, aerosols, land-use changes, water management). They should also quantify changes in river run-off and water quality. Such models would guide regional strategies for adapting to climate change, for preserving and restoring ecosystems and their services, and for conserving biodiversity.

Scientists around the world in multiple disciplines, from climatology to social science, must work together. And local people's needs must be central. Researchers should help communities to understand what is happening to their climate and environment, and enable them to craft strategies for managing risks and adaptation. For example, scientists' assessments of major ice collapses of the Aru glaciers in western Tibet in 2016 enabled the local government to establish a hazard-warning system and to relocate threatened communities.

As the impacts of global warming reverberate around the third pole, science must be at the fore. ■

Jing Gao *is an associate professor at the Institute of Tibetan Plateau Research, Chinese Academy of Sciences (CAS), Beijing; and at the CAS Center for Excellence in Tibetan Plateau Earth Sciences, Beijing, China.* **Tandong Yao** *is a member of the Chinese Academy of Sciences (CAS); a professor at the Institute of Tibetan Plateau Research, CAS, Beijing; and at the CAS Center for Excellence in Tibetan Plateau Earth Sciences, Beijing, China.* **Valérie Masson-Delmotte** *is a senior scientist at the Laboratory for Sciences of Climate and Environment (CEA-CNRS-UVSQ/IPSL), Gif-sur-Yvette, France.* **Hans Christian Steen-Larsen** *is a senior researcher at the Geophysical Institute, University of Bergen; and at the Bjerknes Centre for Climate Research, Bergen, Norway.* **Weicai Wang** *is an associate professor at the Institute of Tibetan Plateau Research, Chinese Academy of Sciences (CAS), Beijing; and at the CAS Center for Excellence in Tibetan Plateau Earth Sciences, Beijing, China.*
e-mail: gaojing@itpcas.ac.cn

1. Immerzeel, W. W., van Beek, L. P. H. & Bierkens, M. F. P. *Science* **328**, 1382–1385 (2010).
2. Yao, T. *et al. Nature Clim. Change* **2**, 663–667 (2012).
3. Farinotti, D. *et al. Nature Geosci.* **8**, 716–715 (2015).
4. Zhang, G. *et al. Geophys. Res. Lett.* **44**, 252–260 (2017).
5. Huss, M. & Hock, R. *Nature Clim. Change* **8**, 135–140 (2018).
6. Roxy, M. K. *et al. Nature Commun.* **6**, 7423 (2015).
7. Lutz, A. F., Immerzeel, W. W., Shrestha, A. B. & Bierkens, M. F. P. *Nature Clim. Change* **4**, 587–592 (2014).
8. Yao, T. *et al. Rev. Geophys.* **51**, 525–548 (2013).
9. Li, W. *Nature Commun.* **9**, 4243 (2018).

© 2019 Springer Nature Limited. All rights reserved.



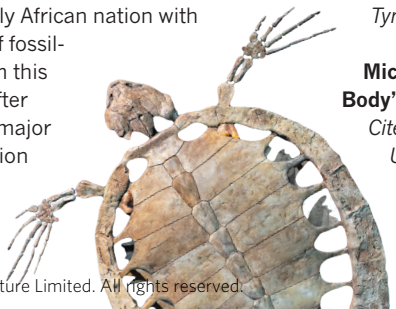
A visitor examines some of Leonardo da Vinci's writing at the Water as Microscope of Nature exhibition at the Uffizi Gallery in Florence, Italy.

Hot tickets 2019

Food and the gut go on show as Angolan 'sea monsters' resurface and Alexander von Humboldt pops into focus. The year is also a feast of anniversaries, from the eclipse proving Albert Einstein right to Leonardo da Vinci's death — and the first footfall on the Moon. **Nicola Jones** reports.

Alexander von Humboldt
Museum Ludwig, Cologne, Germany.
Until 27 January.
Prussian polymath and explorer Alexander von Humboldt's 250th birthday rolls around this September. The 'father of environmentalism' is credited with envisioning geology, ecology and humanity as part of an interconnected web. Less well known is his role in early photography. In 1839, Humboldt was among the first established scientists to embrace the daguerreotype, invented by Louis-Jacques-Mandé Daguerre and Nicéphore Niépce. On show will be photo albums from Humboldt's collection — one a present from British photographic innovator William Henry Fox Talbot; another with some of the first-known photographs of Mexico, Venezuela and Cuba.

Sea Monsters Unearthed: Life in Angola's Ancient Seas
National Museum of Natural History, Washington DC.
Until 2020.
Some 130 million years ago, the supercontinent Gondwana was being ripped apart, forming Africa and South America. The South Atlantic Ocean emerged between them. Today, Angola is a hotspot for tracking the sea's biological record: it is the only African nation with known outcrops of fossil-bearing rocks from this period. In 2005, after decades of war, a major geological expedition reached the region. Projecto PaleoAngola unearthed a new dinosaur species, the long-necked sauropod *Angolatitan adamastor*; a host of sea turtles (**pictured**); and giant marine reptilian plesiosaurs and mosasaurs. Full-scale reconstructions and fossils will be on display at the US National Museum of Natural History. Meanwhile, the museum's David H. Koch Hall of Fossils will open on 8 June with Deep Time, featuring 700 specimens and the return of a *Tyrannosaurus rex* fossil.



SMITHSONIAN

Microbiota: The Inside Story of Our Body's Most Underrated Organ
Cité des Sciences et de l'Industrie, Paris.
Until July.
In 2014, microbiology student Giulia Enders penned a book

M. DEGL' INNOCENTI/UFFIZI GALLERIES

© 2019 Springer Nature Limited. All rights reserved.

Light on Leonardo

The luminary of art and science in the Italian Renaissance, Leonardo da Vinci, died 500 years ago this May. The anniversary will be marked across Europe.

Water as Microscope of Nature: Leonardo da Vinci's Codex Leicester
Uffizi Gallery, Florence, Italy.
Until 20 January.
Here's a chance to gaze at some of Leonardo's spellbinding scientific explorations, not far from his Tuscan birthplace. The Codex Leicester, on loan from Microsoft founder Bill Gates, is a 72-page, mirror-written notebook. It features beautiful images of the movement and erosive capacity of water, an explanation of why fossils are found on mountains, speculation that the Moon's shine is caused by surface water, and more. The codex will also be on display, with two others, in Mind in Motion at the British Library in London, from 7 June to 8 September.

Leonardo da Vinci: A Life in Drawing
Across Britain.
1 February – 6 May.
One hundred and forty-four drawings capturing the range of Leonardo's interests, from painting and music to engineering and botany, will feature across a dozen shows in UK cities from Belfast to Sheffield. Then, from 24 May to 13 October, they will be part of an exhibition featuring 200 drawings (an example **pictured**) at the Queen's Gallery in London (including two apparently blank pages that under ultraviolet light reveal faded studies of hands). From 22 November 2019 to 15 March 2020, 80 of these will grace the Queen's Gallery in Edinburgh.

Leonardo da Vinci and Perpetual Motion: Visualizing Impossible Machines
Peltz Gallery, Birkbeck University of London.
6 February – 12 March.
Featured will be digital reconstructions and 3D printouts of 'perpetual motion' machines designed by Leonardo. Models of perpetual wheels that appear in two of his famous notebooks (the Codex Forster II and the Codex Atlanticus) are included.

Milan and Leonardo
Milan, Italy.
2 May onwards.
This nine-month extravaganza in the artist-scientist's adopted city kicks off with the reopening of the Sala delle Asse in Sforza Castle on 2 May. The nature-inspired murals that Leonardo painted for the Duke of Milan in this room have been under restoration for several years. A virtual tour of what Milan looked like in Leonardo's time is also in the offing.



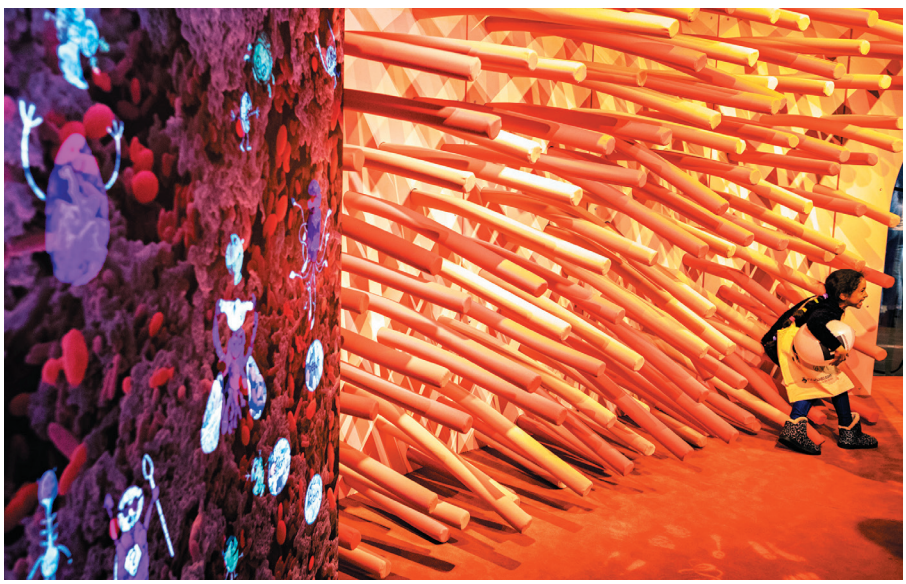
RCT/HER MAJESTY QUEEN ELIZABETH II 2018

Leonardo da Vinci
The Louvre, Paris.
24 October 2019 – 24 February 2020.
At this blockbuster show, drawings (some from Britain's Royal Collection) will be shown along with many of the 17 paintings now attributed to Leonardo. They join his five notable paintings held by the Louvre, including the *Mona Lisa*.

on the digestive system, *Gut* (in her native German, *Darm mit Charme*, or 'Charming Bowels'). The bestseller inspired this exhibition, produced with the French National Institute for Agricultural Research among other bodies, and with Enders's input. The show (**pictured**) echoes the book's lighthearted approach to shadowy topics such as the workings of sphincters inner and outer, and the benefits of squatting for defecation.

By the Light of the Silvery Moon
National Gallery of Art, Washington DC.
28 April – 14 October.
On 20 July 1969, Apollo 11 delivered astronauts to the Moon for the first time. To mark the 50th anniversary, 50 lunar portraits spanning more than a century feature here. On show will be a stereograph of the Moon, taken by English astronomer Warren de la Rue in the 1850s; alongside will be close-up stereographs of the lunar surface by Neil Armstrong and Buzz Aldrin, and iconic photos from early uncrewed orbiters. Also on view will be French astronomer Charles Le Morvan's 1914 photogravures (richly detailed etched copper plates produced from photographic negatives): an attempt to systematically map the visible lunar surface.

PHILIPPE LEVY/EPPDSCIPR



Food
Victoria and Albert Museum, London.
18 May – 17 November.
How we gather, hunt, grow and process food has changed dramatically, from the invention of agriculture some 10,000 years ago to the creation of weight-loss supplements. This exhibition looks at the science and politics of what we eat, from urban farming to synthetic, lab-grown meat fibres and algae-coated spheres of water. New commissions and pieces from the museum's collection will examine how we can make food more sustainable, tasty and ethically just.

Natalia Goncharova
Tate Modern, London.
6 June – 8 September.
In the early twentieth century, Russian avant-garde artist Natalia Goncharova and her partner Mikhail Larionov developed rayonism — an artistic style drawing inspiration from contemporary scientific

© 2019 Springer Nature Limited. All rights reserved.



FRANCO STELLA, FS-HUF-PG/SHF

An artist's impression of the Humboldt Forum in Berlin, opening in late 2019.

breakthroughs. The discoveries of X-rays and radioactivity prompted the artists to represent material reality beyond the reach of the human eye, through fractured shapes and dynamic lines representing light. Descriptions of a mystical 'fourth dimension' by Russian mathematician Peter Ouspensky in the early 1900s resonate in their work, too. Also on show will be Goncharova's futurist body art, and set and clothing designs.

Total Solar Eclipse
La Silla Observatory, Chile.
2 July, 16:39.
This year's total solar eclipse will be visible only from South America and the Southern Pacific Ocean — a fitting moment for La Silla Observatory as it celebrates its first half-century. La Silla marks this portentous alignment with talks, tours and workshops for eclipse-chasers. The year is a big one for astronomy and physics in other ways, too. The International Astronomical Union turns 100, and May sees another centenary: that of astronomer Arthur Eddington's momentous use of a total solar eclipse to secure the first experimental proof of Albert Einstein's general theory of relativity.

Graffiti as Devotion along the Nile
Kelsey Museum of Archaeology, Ann Arbor, Michigan.
23 August 2019 – 5 January 2020.
In Kush, an ancient kingdom in what is now Sudan, many pilgrims left their mark on temples and pyramids starting around 300 BC. Archaeologists examining the graffiti have found inscriptions in ancient Egyptian and Greek describing everything from festivals to diplomatic missions. This show explores hundreds of pieces of pictorial graffiti — including horses, a man with a bow and a giraffe — discovered in a rock-cut temple by Kelsey Museum archaeologists in El Kurru, northern Sudan.

The Future and Arts
Mori Art Museum, Tokyo.
19 November 2019 – 29 March 2020.
Technology is intruding ever further into our everyday lives, from the artificial intelligence that guides smartphone voice activation to blockchains used to 'mine' cryptocurrency. This exhibition asks artists to reflect on what life might be like a few decades in the future — for better or for worse.

Medicine Galleries
Science Museum, London.
Opening in 2019.
At more than 3,000 square metres, these galleries will offer a permanent space for contemplation of the human body and health in London's Science Museum. Some 2,500 artefacts from its archives, and those of the Wellcome Collection across town, will illuminate 500 years of medical history. A 3.5-metre sculpture will greet visitors — the work of Marc Quinn (best known for *Self*, a series of casts of his own head in frozen blood). Called *Self-Conscious Gene*, it portrays the late Canadian artist Rick Genest, who covered his body with tattoos of his skeleton and organs.

The Humboldt Forum
Berlin Palace.
Opening end of 2019.
Berlin's massive, €595-million (US$674-million) museum and event space is set to rival London's British Museum. This 'palace for all' — named after naturalist Alexander von Humboldt and his diplomat brother Wilhelm — will house old and new collections, including those of the Ethnological Museum of Berlin and the Museum of Asian Art. The Humboldt Laboratory will be dedicated to showcasing research from the city's Humboldt University. ■

Screen and stage

Drama, film and television programmes in 2019 will grapple with Ebola, space travel, the disturbing roots of gynaecology and more.

Behind the Sheet
Ensemble Studio Theatre, New York City.
9 January – 3 February.
Controversial nineteenth-century surgeon Marion Sims developed an operation for vesicovaginal fistula, a birth complication that leaves women incontinent. Egregiously, he experimented without anaesthesia on enslaved women of colour in Alabama. They had no power to refuse or consent. Playwright Charly Evon Simpson tackles the story in a drama exploring the experiences of three of them — Anarcha, Lucy and Betsey (**pictured**, Naomi Lorrain, who plays Philomena, the doctor's enslaved assistant).



ENSEMBLE STUDIOS THEATRE

Ad Astra Twenty years after his father disappeared on a mission to Neptune, a US Army Corps engineer (Brad Pitt) sets out to discover what happened, in this blockbuster film directed by James Gray. US release: 24 May.

The Hot Zone In 1989, the US Army and Centers for Disease Control and Prevention spun into action when an Ebola outbreak killed dozens of monkeys in a primate facility just outside Washington DC. The gripping tale of the response, as told in the 1995 book of the same name by Richard Preston, will play out in a television mini-series on National Geographic this year. Juliana Margulies plays real-life virologist Nancy Jaax, who at the time was chief of the pathology division at the US Army Medical Research Institute of Infectious Diseases in Frederick, Maryland. A major twist awaits.

Gemini Man Director Ang Lee presents a sci-fi action-thriller featuring Will Smith as an ageing assassin who must fight a younger clone of himself. US release: 4 October.

© 2019 Springer Nature Limited. All rights reserved.

Correspondence

NIH reviewers on sex-inclusion policy

Since 2016, the US National Institutes of Health (NIH) has required those it funds to consider sex as a biological variable in their experimental design, analyses and reporting of preclinical studies: in other words, they should include female animals equitably, where necessary for rigour. To explore how the policy has been working, we surveyed the scientists who review NIH grant applications — called study-section members — in September 2016 and October 2017. In their responses, we found cause for both commendation and concern.

These reviewers report that an increasing number of investigators are incorporating the policy into their submissions (for details, see N. C. Woitowich and T. K. Woodruff *J. Womens Health* **28**, 9–16; 2019). In 2017, 68% thought that considering sex as a biological variable is important for NIH-funded research, and 58% thought that implementing the policy would improve the rigour and reproducibility of biomedical research. Although study-section members are a subset of all biomedical scientists, their views are an important proxy for the promise of this policy for improving scientific discovery and outcomes.

The quantitative data were positive overall, but female study-section members in the 2017 cohort (the minority) were significantly more likely than men to view the sex-inclusion policy as favourable.

Open-ended comments revealed variability in how policy adherence was judged to affect grant scoring. Some did not consider the policy to be a score-driving factor. Others differed on how it relates to costs and to the overuse of experimental animals. Federal and local dialogue and education should address those concerns.

The swift uptake of the sex-inclusion policy contrasts with the slow progress on the inclusion of women and minorities in NIH-funded clinical research, as stipulated in the 1993 NIH Revitalization Act (S. E. Geller *et al. Acad. Med.* **93**, 630–635; 2018).

Nicole C. Woitowich, Teresa K. Woodruff *Feinberg School of Medicine, Northwestern University, Chicago, Illinois, USA.*
tkw@northwestern.edu

Help islands cope with climate change

Small-island developing states are among the most vulnerable to the effects of climate change. They are fighting rising sea levels and temperatures. To help these states, many of which are members of the commonwealth, the Association of Commonwealth Universities launched the Commonwealth Climate Resilience Network last year. The network's institutions collaborate on mutually beneficial research projects and share best practices for preventing and responding to natural disasters.

These institutions include the University of the West Indies, the University of the South Pacific and Fiji National University. Their research and development draws from information on weather modelling, for example, and guidance on matters such as agricultural technology and big-data collection and analysis. As hubs with local, national and international roles and connections, universities are also crucial for a community's economy in the aftermath of natural disasters.

There are important local initiatives with support from international grants and scholarships. One is the Quake Centre, established in partnership with New Zealand's government and the University of Canterbury as well as several of its industry groups. Another is India's Tata Institute of Social Sciences, which is working with the Kerala government on a long-term institutional response to flooding by using digital systems.

Joanna Newman *Association of Commonwealth Universities, London, UK.*
Joanna.newman@acu.ac.uk

Gene drives: equity demands civility

At the 14th Conference of the Parties of the United Nations Convention on Biological Diversity late last year, I witnessed the rapid deterioration of a crucial discussion. It was on the potential of synthetic biology in environmental conservation. What started as heckling turned into a yelling match of misinformation. Such disruptive behaviour robbed the global community of a rare opportunity to debate gene drives in a meaningful way. Sidelined young scientists, country delegates and others watched in disbelief.

This dangerous breakdown in civil dialogue stems from the potential risks posed by gene-drive technology. In theory, gene drives could restore threatened ecosystems and eliminate vectors of disease. But they could also transform entire species by pushing edited genes through populations of wild plants and animals.

Broad, thoughtful and respectful debate is therefore the only way to abolish scientific and societal blind spots, minimize risks and steer the safe and equitable sharing of any benefits of gene-drive technology. Gene drives are likely to affect environments bound by kinship, cultural identity and life-sustaining resources. It is not enough for the communities in those environments, including historically marginalized peoples, simply to be present at the debating table — their voices must be heard.

Natalie Kofler *Yale Institute for Biospheric Studies, New Haven, Connecticut, USA.*
natalie.kofler@yale.edu

Rwanda takes up science-policy baton

Rwanda is now one of the growing number of developing countries that are encouraging scientists to help shape evidence-based government policy (see also *Nature* **560**, 671–673; 2018). Such efforts (see, for example, go.nature.com/2eae4kpn) have boosted health gains in the past couple of decades. World Bank data show that life expectancy in Rwanda rose from 29 to 68 years between 1994 and 2015, and mortality during childbirth fell by 70% over the same period.

The Rwanda Biomedical Center in Kigali City, for instance, works in partnership with researchers across the world in the Demand-Driven Evaluations for Decisions programme to analyse service statistics from Health Management Information Systems and to answer policymakers' research questions. Data collected from different districts can be used to assess the impact of expanding community-based malaria treatment during times of resurgence, for example. Health professionals can then use the outcomes of policy interventions that are based on such evidence to improve local clinical practice.

Clarisse Musanabaganwa*
Rwanda Biomedical Center, Kigali, Rwanda.
**On behalf of 5 correspondents (see go.nature.com/2sycg5j for details).*
clarisse.musanabaganwa@rbc.gov.rw

CONTRIBUTIONS

Correspondence may be submitted to **correspondence@nature.com** after consulting the author guidelines and section policies at **http://go.nature.com/cmchno**.

© 2019 Springer Nature Limited. All rights reserved.

CHEMICAL BIOLOGY

Engineered enzymes set a trap

Many enzymatic processes involve a mechanism in which reaction intermediates are covalently attached to the enzyme's active site. A strategy has been devised that enables mimics of these intermediates to be visualized. SEE LETTER P.112

ANDREW M. GULICK

Enzyme structure and function are routinely studied by altering the DNA that encodes the enzyme, thus replacing specific amino-acid residues in the enzyme with other residues. Unfortunately, the substitute residues are generally limited to those derived from amino acids that exist in nature (canonical amino acids). One strategy for overcoming this constraint is to use genetic-code expansion to engineer the cell's protein-synthesis machinery. In this approach, enzymes called amino acyl-tRNA synthetases are designed to load non-canonical amino acids (ncAAs) onto transfer RNAs, so that the ncAAs can be incorporated into proteins by the cell's ribosome apparatus[1]. On page 112, Huguenin-Dezot *et al.*[2] report a method for studying enzyme mechanisms that uses genetic-code expansion in a new way. The authors demonstrate the utility of this approach using an enzyme that produces an antibiotic.

Many enzymes form covalent chemical bonds to the reacting molecule as part of their strategy for catalysing reactions. In one common class of enzyme, this process generates acyl-enzyme intermediates in which a covalent bond has connected a nucleophilic group (most commonly a serine or cysteine amino-acid residue) in the enzyme's active site to a carboxylic acid group (COOH) in the reactant; this produces either an ester or a thioester linkage from a serine or cysteine residue, respectively (Fig. 1a). In a second step, the covalent bond is broken through a hydrolysis reaction, releasing the product of the enzymatic reaction and restoring the reactive nucleophile for a second round of catalysis.

Visualization of the acyl-enzyme intermediates by X-ray crystallography is often challenging because spontaneous hydrolysis can occur. Huguenin-Dezot and colleagues have developed a system that allows the nucleophilic amino-acid residue to be replaced with an ncAA residue called 2,3-diaminopropionic acid (DAP). In effect, this switches the hydroxyl group (OH) of serine — or the thiol group (SH) of cysteine — for an amine group (NH_2). The acyl-enzyme intermediate that forms from the DAP residue is resistant to hydrolysis, and is therefore suitable for structural characterization (Fig. 1b).

The chemical similarities between DAP, serine and cysteine amino-acid residues made it difficult for the authors to engineer an amino acyl-tRNA synthetase that selectively loads only the ncAA. They therefore attached a 'masking' group to the side chain of DAP so that the modified ncAA could be distinguished from the canonical amino acids. Once the modified DAP has been incorporated into an enzyme, the masking group is removed by exposing the protein to ultraviolet light.

Huguenin-Dezot *et al.* tested five DAP analogues that had different masking groups, and found one that accumulates in protein-producing bacteria at sufficient concentrations to enable its incorporation into proteins. The researchers then engineered an amino acyl-tRNA synthetase to function with this modified DAP, and demonstrated that the DAP analogue could be incorporated, and subsequently unmasked, in a test protein (green fluorescent protein). As a second test case, they incorporated DAP into the active site of a cysteine protease enzyme, and showed that the inserted DAP indeed forms an acyl-enzyme intermediate that is resistant to hydrolysis.

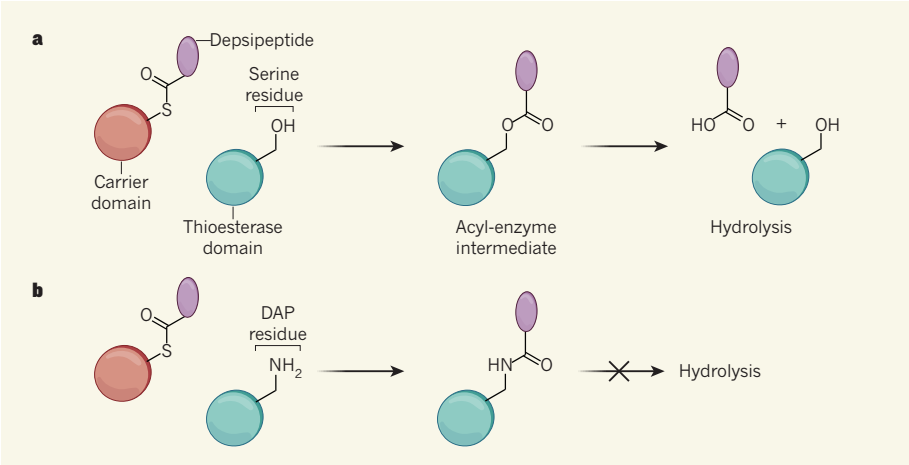


Figure 1 | Stabilizing reaction intermediates of enzyme reactions. Enzymes known as non-ribosomal peptide synthetases assemble peptides or depsipeptides (peptides that contain one or more ester linkages). These enzymes use carrier domains to shuttle the growing product to catalytic domains for modification and extension. **a**, The bound product is released from the final carrier domain through a reaction catalysed by the thioesterase domain. A nucleophilic amino-acid residue (often, a serine residue, the side chain of which is shown here) in the thioesterase domain reacts with the linkage that attaches the depsipeptide to the carrier domain. This produces an acyl-enzyme intermediate, which is subsequently hydrolysed to release the depsipeptide. **b**, Huguenin-Dezot *et al.*[2] report a method in which the nucleophilic residue is substituted by the 2,3-diaminopropionic acid (DAP) residue — which here replaces the hydroxyl (OH) group of serine with an amine (NH_2) group. The resulting acyl-enzyme intermediate resists hydrolysis, allowing it to be visualized using X-ray crystallography.

The authors went on to use DAP to study the antibiotic-producing enzyme valinomycin synthase. This enzyme belongs to a family of proteins known as non-ribosomal peptide synthetases (NRPSs), which construct peptides or depsipeptides (peptides that contain one or more ester linkages). NRPSs have a fascinating 'assembly-line' architecture: the growing peptide is attached to carrier domains that shuttle the intermediates to neighbouring catalytic domains, where the peptide is modified and extended[3].

This biosynthetic strategy requires that the bound product is ultimately released from the final carrier domain, freeing up that domain to conduct additional synthetic cycles. The release occurs through a reaction carried out by the thioesterase domain. The thioesterase domain catalyses this reaction by forming a covalent bond to the peptide to generate an initial acyl-enzyme intermediate, which then either is hydrolysed or undergoes a macrocyclization reaction to release the peptide; macrocyclization reactions are those in which the ends of a long peptide chain join

© 2019 Springer Nature Limited. All rights reserved.

together to form a large, ring-shaped molecule.

Valinomycin consists of three copies of a short depsipeptide (a tetradepsipeptide), which are joined together by the thioesterase domain of valinomycin synthase to form a ring[4]. Using the wild-type enzyme, Huguenin-Dezot and colleagues first demonstrated that the thioesterase uses a 'retrograde' reaction mechanism to join the three tetradepsipeptides together. After the first tetradepsipeptide has been loaded onto the nucleophilic residue in the thioesterase domain, the second tetradepsipeptide arrives at the thioesterase active site, attached to the carrier domain. The first tetradepsipeptide is then transferred to, and reacts with, the second one, temporarily placing the resulting octadepsipeptide product on the carrier domain (rather than the second tetradepsipeptide being transferred to the first one on the thioesterase domain). The octadepsipeptide then migrates to the thioesterase domain, and the third tetradepsipeptide is brought to the active site to repeat the reaction cycle. This generates a dodecadepsipeptide that finally undergoes macrocyclization to form valinomycin.

The authors next incorporated DAP into the thioesterase domain to produce a stable acyl-enzyme intermediate that could be visualized using X-ray crystallography. The full structure of the dodecadepsipeptide could not be clearly visualized, but the residues that could be seen suggest that the active site of the thioesterase guides macrocyclization, partly by excluding water to minimize hydrolysis, and partly by directing the peptide's terminals towards each other for the reaction. The X-ray structure also shows that part of the thioesterase domain known as the lid region undergoes a substantial conformational change to orient the depsipeptide correctly for reaction.

When the serine or cysteine residue of an enzyme is replaced by DAP, the corresponding ester or thioester linkage in the acyl-enzyme intermediate is replaced by an amide linkage. It should be noted that this is not a perfect mimic of the natural linkage. The very feature of the amide that makes it resistant to hydrolysis — the fact that it behaves in part like a double bond — also constrains it to a planar geometry. Such a restriction does not apply for ester and thioester linkages. This is a minor issue, however, and does not diminish the value of visualizing the amide analogue of the natural acyl-enzyme intermediate.

Huguenin-Dezot and co-workers' method could be used to study various enzymatic systems, including enzymes of the ubiquitination pathway[5] (which marks proteins for degradation), or any of the many types of hydrolase enzyme used in diverse bond-cleaving processes[6,7]. Previous efforts to visualize the structures of acyl-enzyme intermediates have often used molecules that react irreversibly with the nucleophilic amino-acid residues in active sites[8], to produce stable analogues of acyl-enzyme intermediates — but these might not mimic the ester (or thioester) linkage as closely as do Huguenin-Dezot and colleagues' amide analogues.

Furthermore, the new work could complement the use of ncAAs that bear reactive groups, which can be incorporated into large, multidomain enzymes to form covalent cross-links between proteins or protein domains to trap functional states for structural studies. This has previously been done for polyketide synthase enzymes (which are closely related to NRPSs) at non-catalytic residues positioned at domain interfaces[9]. Huguenin-Dezot and co-workers' introduction of ncAA residues at positions that are responsible both for chemistry and for recognizing enzyme substrates or partner proteins, perhaps used in combination with chemical analogues of substrates and intermediates[8], might uncover previously unknown structural features of biosynthetic enzymes.

Finally, DAP residues could potentially be used in a complementary method to existing approaches for profiling enzyme function[10]: they could trap molecules from pools of metabolites in the active sites of uncharacterized hydrolases, and the bound molecule could be then identified using methods based on mass spectrometry to define the enzyme's preferred substrate. With such a wide variety of applications, Huguenin-Dezot and colleagues' method is a valuable addition to the protein chemist's toolbox that will advance the study of many important enzymes. ■

Andrew M. Gulick *is in the Department of Structural Biology, Jacobs School of Medicine and Biomedical Sciences, University at Buffalo, Buffalo, New York 14203, USA.*
e-mail: amgulick@buffalo.edu

1. Wang, L., Brock, A., Herberich, B. & Schultz, P. G. *Science* **292**, 498–500 (2001).
2. Huguenin-Dezot, N. *et al*. *Nature* **565**, 112–117 (2019).
3. Payne, J. A., Schoppet, M., Hansen, M. H. & Cryle, M. J. *Mol. Biosyst.* **13**, 9–22 (2016).
4. Cheng, Y. Q. *ChemBioChem*. **7**, 471–477 (2006).
5. Olsen, S. K., Capili, A. D., Lu, X., Tan, D. S. & Lima, C. D. *Nature* **463**, 906–912 (2010).
6. Bachovchin, D. A. & Cravatt, B. F. *Nature Rev. Drug Discov.* **11**, 52–68 (2012).
7. Lentz, C. S. *et al. Nature Chem. Biol.* **14**, 609–617 (2018).
8. Gulick, A. M. & Aldrich C. C. *Nat. Prod. Rep.* **35**, 1156–1184 (2018).
9. Ye, Z. & Williams, G. J. *Biochemistry* **53**, 7494–7502 (2014).
10. Backus, K. M. *et al*. *Nature* **534**, 570–574 (2016).

This article was published online on 12 December 2018.

INFLUENZA

All for one and one for all to fight flu

Antibodies have been engineered to recognize diverse strains of influenza, including both the A and B types of virus that cause human epidemics. Are we moving closer to achieving 'universal' protection against all flu strains?

GARY J. NABEL & JOHN W. SHIVER

The First World War, one of the world's deadliest conflicts, claimed approximately 20 million lives. But in the year that it ended, an even deadlier calamity swept the globe. The 1918 influenza pandemic is estimated to have killed between 50 million and 100 million people[1]. Within months, a simple virus took a greater toll on human life than had the brutal four-year war. Although flu vaccines now save countless lives and have undoubtedly helped to avert other worldwide pandemics, their manufacture must vary annually to match the current circulating viral strains. Flu continues to pose a threat to human health, and there is a pressing need to develop countermeasures that can confer broad protection against the different flu strains. Moreover, vaccines are usually less effective at providing protection for children and for older individuals than they are for the rest of the population[2].

Writing in *Science*, Laursen *et al.*[3] report that they have engineered antibodies that protect against diverse flu viruses in mice, and that notably provide protection against most viral strains from the two major types of virus that cause human disease: influenza A and influenza B. Obtaining such broad protection has been a challenge because both influenza A and influenza B are composed of diverse strains, and developing 'universal' protection has been an elusive goal. If the authors' approach could be adapted for effective use in humans, it might help to prevent or contain the spread of new and evolving flu infections worldwide.

During the 1918 pandemic, the cause of the disease was unknown. Had a vaccine been available, it would probably have limited the global catastrophe. However, developing an effective flu vaccine is not straightforward because flu viruses can mutate rapidly[4]. The high level of mutation results in continuous variation in two key viral proteins over time. One of these is haemagglutinin, which is located on the surface of the virus (Fig. 1) and

© 2019 Springer Nature Limited. All rights reserved.

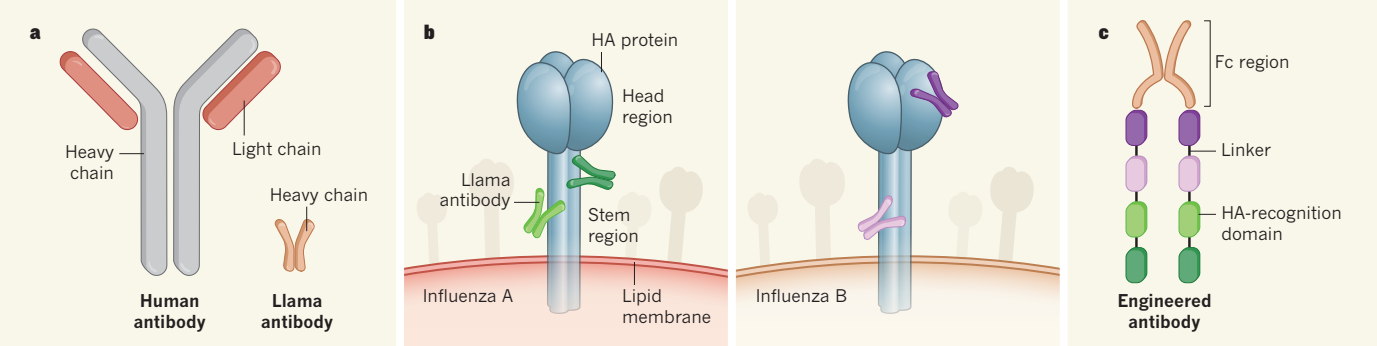


Figure 1 | Engineered antibodies target diverse strains of influenza virus. Laursen *et al.*[3] report the development of antibodies that can provide broad protection against flu strains when tested in mice. **a**, The authors engineered antibodies based on antibodies from llamas (*Lama glama*), which are smaller than human antibodies. Llama antibodies, like human antibodies, contain regions known as heavy chains, but lack structures called light chains. **b**, The authors assessed llama antibodies that target the protein haemagglutinin (HA), which is found on the surface of the flu virus. *In vitro* analyses enabled the identification of antibodies that provide potent protection against the virus, and the authors isolated antibodies that targeted the two major groups of the virus: influenza A (green antibodies)and influenza B (purple antibodies). Structural analysis indicated whether the antibodies bound to the 'stem' or 'head' region of HA. **c**, The authors engineered antibodies containing such HA-recognition domains from llama antibodies, joined by linker regions (black). The antibodies also included an Fc region to aid interaction with immune cells. These antibodies provided protection against selected influenza A and B strains when tested in mice.

recognizes a molecule on host cells that serves as a receptor for viral attachment and entry.

Haemagglutinin also associates with a viral protein called neuraminidase. There are 18 distinct subtypes of haemagglutinin and 11 versions of neuraminidase. These two proteins form the basis for how flu strains are named. For example, the designation H1N1 indicates that a flu virus has haemagglutinin subtype 1 and neuraminidase subtype 1.

A breakthrough in attempts to offer protection against diverse flu strains came with the identification of antibodies, termed broadly neutralizing antibodies, that can bind to a highly evolutionarily conserved and invariant structure in a region of haemagglutinin termed the stem[5,6]. Such antibodies battle the flu virus by binding haemagglutinin and inhibiting the virus's ability to enter cells. They can also boost an antiviral response, for example by engaging immune cells that promote the killing of virus-infected cells. However, these antibodies usually do not recognize all flu viruses. For instance, broadly neutralizing antibodies that recognize haemagglutinin of one major genetic subgroup of influenza A, group 1, typically do not react against the second group, group 2, and also do not recognize influenza B (ref. 7).

To try to target influenza A and influenza B viruses, Laursen and colleagues had the idea of engineering an antibody by 'stitching' together influenza-recognition domains from different antibodies that bind to evolutionarily conserved regions of haemagglutinin, especially in the stem region of this protein. The authors vaccinated llamas (*Lama glama*) with a flu vaccine or haemagglutinin proteins, and used *in vitro* tests to identify resulting antibodies that had the greatest potency and breadth of neutralization against diverse flu viruses. They found that specific combinations of these antibodies could target nearly all of the flu-virus strains tested. Llama antibodies have a simpler structure and are smaller than human antibodies, and therefore aid an engineering approach that seeks to combine protein regions from more than one antibody.

By engineering antibodies in which several influenza-recognizing regions were connected by protein linkers, the authors were able to create antibodies that targeted multiple viruses. And fusing such structures to an antibody structure called the Fc region enabled such chimaeric proteins to interact with and activate immune cells.

When mice received either engineered antibodies or the gene encoding such an antibody — delivered by means of an adeno-associated virus (AAV) into cells of the nasal passage — they were protected against a flu virus that would usually have been lethal. The gene-delivery approach ensured production of the antibody for weeks to months, providing sustained protection without the need for multiple rounds of antibody injections over time.

Whether this approach could be used to prevent flu in humans is uncertain. Mice do not serve as optimal models for investigating human influenza because the receptor used by viral strains to infect mouse cells is a different version from that needed for cellular entry into human cells. In addition, the patterns of tissue infection and virus in the bloodstream often differ between mice and humans[8]. Protection in mice can involve a pathway that is mediated by a receptor protein called FcγR-III on immune cells that recognizes antibodies bound to targets[9], but whether this type of immune mechanism has relevance for humans is unknown. Antibodies that target the stem region of haemagglutinin have so far failed to alleviate symptoms in humans who are already infected, and the ability of these antibodies to prevent infection is being tested in clinical trials[10].

Another worry regarding this approach in humans is whether an immune response might be triggered against the non-human antibodies. Although an engineered llama antibody has been approved for clinical use to treat a blood-clotting condition[11], whether an immune response is generated against the anti-flu multi-domain antibodies will become clear only with clinical testing. Llama antibodies can be 'humanized' (engineered to closely resemble the related domains of human antibodies), yet the efficacy of such modifications would need to be evaluated in humans.

Also of concern is the use of AAV, because there are limitations to achieving sufficient and sustained levels of gene expression when using this virus in gene-therapy treatments[12]. Other safety and regulatory concerns regarding AAVs relate to their use to drive continuous gene expression, because this raises the possibility that complexes of human antibodies bound to the engineered antibodies might form over time. That said, certain groups, such as older people, might especially benefit from engineered antibodies, given the high mortality rates from flu in such individuals, and the fact that their immune responses tend to be less robust than those of younger adults.

The expression of engineered antibodies through gene-delivery approaches might offer a way of preventing or treating diverse types of infectious disease. Moreover, the outcomes of such treatments might help to confirm useful targets for the development of antiviral drugs or vaccines. For example, if broadly neutralizing antibodies that target the stem region of haemagglutinin can prevent flu infection *in vivo* in humans, it would encourage efforts to generate such antibodies by vaccination approaches. Antibodies targeting the stem region of haemagglutinin have previously been generated using structure-based approaches for vaccine design, and have shown promise in preclinical tests using animal models[13–15].

Laursen and colleagues' approach of generating an antibody that can target more than one site is reminiscent of earlier work in which an antibody was developed[16] from broadly neutralizing antibodies to target three independent sites on the HIV virus. Such antibodies can neutralize more than 99% of circulating HIV strains. This antibody blocked infection from viruses that were unaffected if single antibody

© 2019 Springer Nature Limited. All rights reserved.

components of the trispecific antibody were used. The era of multi-specific target engagement by engineered antibodies has begun, and might lead to new countermeasures to protect human health. ■

Gary J. Nabel *is at Sanofi, Cambridge, Massachusetts 02139, USA.* **John W. Shiver** *is at Sanofi Pasteur, Swiftwater, Pennsylvania 18370, USA.*
e-mail: gary.nabel@sanofi.com

1. Morens, D. M. & Fauci, A. S. *J. Infect. Dis.* **195**, 1018–1028 (2007).
2. Paules, C. & Subbarao, K. *Lancet* **390**, 697–708 (2017).
3. Laursen, N. S. *et al. Science* **362**, 598–602 (2018).
4. Taubenberger, J. K. & Kash, J. C. *Cell Host Microbe* **7**, 440–451 (2010).
5. Ekiert, D. C. *et al. Science* **324**, 246–251 (2009).
6. Sui, J. *et al. Nature Struct. Mol. Biol.* **16**, 265–273 (2009).
7. Corti, D. *et al. Science* **333**, 850–856 (2011).
8. Margine, I. & Krammer, F. *Pathogens* **3**, 845–874 (2014).
9. Jegaskanda, S., Reading, P. C. & Kent, S. J. *J. Immunol.* **193**, 469–475 (2014).
10. Nachbagauer, R. & Krammer, F. *Clin. Microbiol. Infect.* **23**, 222–228 (2017).
11. Peyvandi, F. *et al. N. Engl. J. Med.* **374**, 511–522 (2016).
12. Borsotti, C. & Follenzi, A. *Expert Rev. Clin. Immunol.* **14**, 1013–1019 (2018).
13. Hai, R. *et al. J. Virol.* **86**, 5774–5781 (2012).
14. Impagliazzo, A. *et al. Science* **349**, 1301–1306 (2015).
15. Yassine, H. M. *et al. Nature Med.* **21**, 1065–1070 (2015).
16. Xu, L. *et al. Science* **358**, 85–90 (2017).

The authors declare competing financial interests.
See go.nature.com/2zm2zhu for details.

This article was published online on 10 December 2018.

GLACIOLOGY

Greenland's subglacial methane released

Methane produced in sediments beneath the Greenland Ice Sheet is released to the atmosphere by meltwater in the summer. This suggests that glacial melt could be an important global source of this greenhouse gas. SEE LETTER P.73

LAUREN C. ANDREWS

Sediments beneath glaciers and ice sheets harbour carbon reserves that, under certain conditions, can be converted to methane, a potent greenhouse gas. However, the formation and release of such methane is an unquantified component of the Arctic methane budget. On page 73, Lamarche-Gagnon *et al.*[1] present direct measurements of dissolved methane in water discharged from a land-terminating glacier of the Greenland Ice Sheet during the summer. This water, which is known as proglacial discharge (Fig. 1), was supersaturated with methane, and the amount of methane released to the atmosphere from this discharge rivals that from other terrestrial rivers. The findings suggest that the form and evolution of subglacial hydrological systems contribute to the control of the Arctic methane cycle.

Atmospheric methane concentrations varied substantially in the past, and it has been hypothesized[2] that large reserves of methane can form and be trapped under ice sheets and glaciers when there is a favourable combination of carbon-rich sediments, high subglacial pressures, oxygen-poor conditions and low temperatures. Rapid release of this methane during glacial retreat might trigger rapid warming[3], but whether large-scale release of such glacial methane could occur in the future is disputed[4].

Field observations provide equivocal evidence for whether subglacial sediments act as a source or sink of methane. Ice-core drilling operations in West Antarctica detected methane-producing microbes in proglacial and subglacial sediments[5], but analyses of Antarctic subglacial lake sediments[6] and proglacial sediments[7] indicate that bacterial oxidation consumes almost all the methane produced, preventing its release to the atmosphere. This bacterial methane cycling suggests that the subglacial hydrological system could serve as a methane sink.

Lamarche-Gagnon *et al.* show that subglacial microbial oxidation of methane is not sufficient to mitigate the gas's release to the atmosphere within a well-characterized subglacial catchment (an area within which subglacial water collects and drains out of a common outlet) in Greenland. Subglacial sediments can therefore act as a local source of methane, corroborating the results of other recent studies of subglacial methane[8,9]. Lamarche-Gagnon *et al.* go further by demonstrating that the continuous flux of methane from the Greenland subglacial environment varies with the efficiency of subglacial meltwater drainage.

Near the margins of the Greenland Ice Sheet, the glacial hydrological system seems to be well suited to exporting methane. During winter, some meltwater from previous summers is stored in the inactive subglacial hydrological system[10]. Lamarche-Gagnon *et al.* hypothesize that this winter storage allows meltwater to become enriched with methane through interaction with sediments in an oxygen-free environment.

In the spring, the subglacial hydrological system reactivates when it is flooded by the drainage of surface meltwater through crevasses and moulins (vertical conduits that allow meltwater to flow from the glacier surface to the bed beneath the ice sheet). This flooding causes an increase in ice motion[11], and Lamarche-Gagnon and colleagues show that this flushes the methane-enriched, subglacial water to the ice margin. The authors observed multiple distinct flushing events after the activation of the subglacial hydrological system, suggesting that various types of meltwater pulse — drainage of lakes on the ice surface through moulins, rapid surface melting during warm spells, and upstream expansion of active subglacial water systems — can liberate subglacial methane.

The observations indicate that methane concentrations tend to peak following sediment-rich proglacial discharge. The slight delay between peak methane export and proglacial discharge suggests that drainage of methane-rich water occurs just after the lateral flow of water that is associated with the greatest ice motion, and during times when differences in water pressure allow the drainage of normally isolated regions of the ice-sheet bed into enlarged subglacial channels[12,13].



JEREMY HARBECK/NASA

Figure 1 | Sediment-rich meltwater released from Russell Glacier, Greenland. Lamarche-Gagnon *et al.*[1] report that such discharge contains concentrations of methane equal to those of many terrestrial rivers.

© 2019 Springer Nature Limited. All rights reserved.

Lamarche-Gagnon *et al.* posit that the formation and growth of subglacial channels permits the rapid evacuation of stored methane-rich meltwater, limiting the amount of time that it is exposed to the oxygen-rich subglacial hydrological system in which bacterial oxidation occurs.

The effect on the atmosphere of the export of subglacially produced methane — and of any future growth in this methane release associated with ice-sheet retreat — will depend on several as-yet-unconstrained factors. The potential increase of methane production and export in Greenland might be limited by the area of liquid water that can form at the ice-sheet bed[14]. The extent of carbon-rich sediments beneath ice sheets and glaciers is also unknown, particularly in Greenland, where both sediments[15] and hard beds[16] have been observed.

Any increase in the proglacial methane flux from either Greenland or Antarctica will probably require a long-term expansion of subglacial hydrological systems that can efficiently evacuate stored meltwater. In Greenland, efficient subglacial drainage typically extends about 40 kilometres from the ice-sheet edge. Surface meltwater production in Greenland will continue to expand[17], but the surface and basal topography of the ice sheet might limit the extent of efficient subglacial drainage[18], and the nature of the ice flow could limit surface-to-bed connections[19].

Antarctica has extensive regions of subglacial sediments and liquid water. Increased surface meltwater and surface-to-bed connections in the future might encourage more-efficient subglacial drainage in regions where methane is produced and stored. However, any increase in subglacial methane mobilization could be mitigated if water flow is slow or if subglacial basins are large, thus allowing more-complete bacterial oxidation of methane to occur. In such scenarios, subglacial methane export might be limited to regions near the ice terminus.

Lamarche-Gagnon and colleagues' study provides an example of how our planet's icy domains can interact with the surrounding Earth system in unexpected and potentially important ways. Modelling and observational studies that characterize the ability of subglacial sediments to convert and store methane, and the ability of the subglacial hydrological system to export this methane to the atmosphere, will be key steps towards improving our knowledge of the sources and sinks of Arctic methane — and better constraining estimates of their future changes. ■

Lauren C. Andrews *is in the Global Modeling and Assimilation Office, NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA.*
e-mail: lauren.c.andrews@nasa.gov

1. Lamarche-Gagnon, G. *et al. Nature* **565**, 73–77 (2019).
2. Wadham, J. L. *et al. Nature* **488**, 633 (2012).
3. Weitemeyer, K. A. & Buffett, B. A. *Glob. Planet. Change* **53**, 176–187 (2006).
4. Petrenko, V. V. *et al. Nature* **548**, 443–446 (2017).
5. Christner, B. C. *et al. Nature* **512**, 310–313 (2014).
6. Michaud, A. B. *et al. Nature Geosci.* **10**, 582 (2017).
7. Dieser, M. *et al. ISME J.* **8**, 2305–2316 (2014).
8. Christiansen, J. R. & Jørgensen, C. J. *Sci. Rep.* **8**, 16623 (2018).
9. Burns, R. *et al. Sci. Rep.* **8**, 17118 (2018).
10. Chu, W. *et al. Geophys. Res. Lett.* **43**, 12484–12492 (2016).
11. Hoffman, M. J., Catania, G. A., Neumann, T. A., Andrews, L. C. & Rumrill, J. A. *J. Geophys. Res.* **116**, F04035 (2011).
12. Cowton, T., Nienow, P., Bartholomew, I. & Mair, D. *J. Glaciol.* **62**, 451–466 (2016).
13. Hoffman, M. J. *et al. Nature Commun.* **7**, 13903 (2016).
14. MacGregor, J. A. *et al. J. Geophys. Res. Earth Surf.* **121**, 2015JF003803 (2016).
15. Walter, F., Chaput, J. & Lüthi, M. P. *Geology* **42**, 487–490 (2014).
16. Harper, J. T., Humphrey, N. F., Meierbachtol, T. W., Graly, J. A. & Fischer, U. H. *J. Geophys. Res. Earth Surf.* **122**, 2017JF004201 (2017).
17. Leeson, A. A. *et al. Nature Clim. Change* **5**, 51–55 (2015).
18. Meierbachtol, T. W., Harper, J. & Humphrey, N. *Science* **341**, 777–779 (2013).
19. Poinar, K. *et al. Geophys. Res. Lett.* **42**, 2015GL063192 (2015).

CONDENSED-MATTER PHYSICS

Topological properties controlled by light

In materials called Weyl semimetals, electrons form structures that have distinct topological properties. The discovery of an ultrafast switch between two of these structures could have many practical applications. SEE LETTER P.61

YOUNG-WOO SON

When electrons flow through arrays of atoms in certain solids, they behave as quantum-mechanical particles that have extremely high speeds. Graphene — an atomically thin sheet of carbon — is an example of a material in which such behaviour occurs in two dimensions[1]. In the past decade, there have been worldwide efforts to study solids called Weyl semimetals, which exhibit similar intriguing properties of electrons in three dimensions[2]. In Weyl semimetals, electrons form structures that have peculiar topological properties, resulting in many fascinating characteristics of matter[2]. On-demand control of the electronic properties of a Weyl semimetal would therefore enable ultrafast manipulation of the material's properties. On page 61, Sie *et al.*[3] report that terahertz-frequency light can provide such control in a particular Weyl semimetal.

In the late 1920s, the physicist Paul Dirac discovered an equation that governs the behaviour of relativistic (high-speed) particles, and that combines quantum mechanics and Einstein's special theory of relativity[4]. Following this monumental work, the mathematician and theoretical physicist Hermann Weyl suggested a simplified version of Dirac's equation[5], describing massless particles — known as Weyl fermions — that have a chirality (handedness) of −1 or +1. In a Weyl semimetal, the dynamics of low-energy electrons are governed by Weyl's equation.

The quantum state of an electron is characterized by the particle's energy, momentum and spin (intrinsic angular momentum). In a solid, these quantum states are dictated by symmetries of the material's atomic lattice. Under time-reversal symmetry, the physical properties of a material are unchanged when the direction of time is reversed. Under inversion symmetry, the physical properties are retained when the spatial coordinates are flipped. If both of these symmetries are preserved, there are always two quantum states of electrons that have the same energy and momentum.

However, if one of these symmetries is broken, it is still possible to have quantum states of equal energy and momentum at particular points in phase space — the space of all possible energy and momentum values[2,6]. In the vicinity of these points, electrons are described by Weyl's equation. As a result, the elementary excitations of electrons in such solids behave as Weyl fermions, and the associated chiralities can be assigned to states near the points. By analogy with particles of opposite electric charge, states of opposite chirality can be produced in pairs, and annihilate each other when they meet.

Among the handful of Weyl semimetals that have been discovered, molybdenum ditelluride and tungsten ditelluride ($MoTe_2$ and WTe_2, respectively) are of particular scientific interest[2,7]. These compounds contain 2D structures that stack through a weak attractive force — the van der Waals interaction — to form layered 3D crystals. Depending on the stacking geometry, different crystal symmetries can be realized.

It has been known for more than four decades that, as temperature increases, $MoTe_2$ changes from one crystal structure (orthorhombic) to another (monoclinic), whereas WTe_2 does not[8]. Inversion symmetry is broken in the orthorhombic structure, so that the associated compounds can exhibit the Weyl-semimetal phase[2,7] (Fig. 1). By contrast,

© 2019 Springer Nature Limited. All rights reserved.

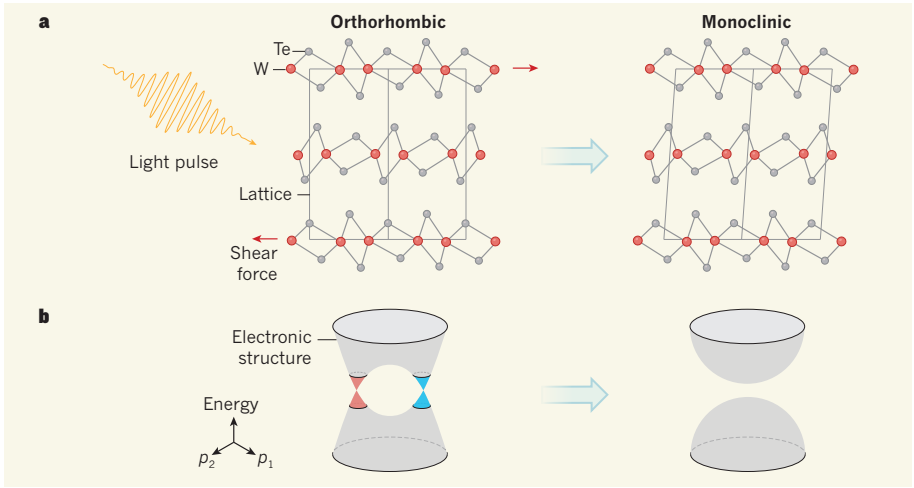


Figure 1 | Switching mechanism in a Weyl semimetal. a, Tungsten ditelluride (WTe_2) is an example of a material known as a Weyl semimetal. Sie *et al.*[3] report a transition in WTe_2 between two crystal structures: orthorhombic and monoclinic. The authors show that terahertz-frequency light pulses can transform the orthorhombic structure into the monoclinic one by altering the material's atomic lattice in a similar way to shear forces — pairs of equal and opposite forces that act on the top and bottom layers of the material. **b**, The dynamics of electrons in a solid can be described by a particular structure in phase space (the space of all possible energy and momentum values). Shown here are highly simplified electronic structures, in which only two components of momentum (p_1 and p_2) are considered. In orthorhombic WTe_2, electrons can behave as massless particles called Weyl fermions that have a chirality (handedness) of −1 (red) or +1 (blue). In monoclinic WTe_2, these states of opposite chirality can annihilate each other.

inversion symmetry is preserved in the monoclinic structure, and the states of opposite chirality can annihilate each other. The two crystal structures have almost the same atomic lattice, except that the monoclinic one is tilted by about 4° with respect to the out-of-plane direction of the orthorhombic one.

Owing to the weak attractive force between the layers of the $MoTe_2$ and WTe_2 compounds, each layer can slide easily, unlike in ordinary materials. As a result, shear forces — pairs of equal and opposite forces that act on the top and bottom layers — can deform the orthorhombic structure into the monoclinic structure, and therefore the Weyl-semimetal phase into a normal phase. Applying such forces in a mechanical way might either permanently alter the atomic lattices or be impossible. A theoretical study suggested that the crystal symmetries of these structures could instead be switched using charge doping, whereby electrons are added to or subtracted from a material[9]. The study indicated that this method might provide a controllable way to switch between the different topological phases.

Sie and colleagues' work is probably the first to demonstrate a dynamic transition between two crystal structures that have distinct topological phases. Previous studies have reported similar topological transitions, but these studies used static mechanical controls that cannot easily switch between the different phases[10,11]. Sie *et al.* found that light pulses at terahertz (THz) frequencies could cause the orthorhombic structure to become unstable by exciting electrons. This could induce the structural transition of WTe_2 from orthorhombic to monoclinic, as if charge doping had been applied to the sample. The authors analysed the crystal structures using a technique known as relativistic ultrafast electron diffraction. They corroborated their measurements using a method called time-resolved second-harmonic generation, which is quite sensitive to the inversion symmetry of crystals.

All the authors' measurements clearly indicate that the crystal structure of WTe_2 has inversion symmetry after the light pulses have been applied, and the switching between structures occurs at THz frequencies — although recovery of the original structure takes much longer. Because the absence of inversion symmetry is a key characteristic of the Weyl-semimetal phase in orthorhombic WTe_2, the observation of this switch of symmetries provides strong indirect evidence of the topological transition. Sie and colleagues have therefore discovered a dynamical way to control the topological properties of Weyl semimetals that could open up many applications, because the existence of Weyl fermions can substantially alter the behaviour of these materials[2].

Further studies are needed to realize the full potential of the authors' switching mechanism. Because the structural transitions in $MoTe_2$ and WTe_2 are closely related to topological changes[9], combined electrical and optical measurements would not only conclusively determine the topological transitions, but also provide a way to study topology-related transport phenomena in these solids[2]. The microscopic description of how THz-frequency light pulses affect the electronic and structural properties of WTe_2 is also required to understand the observed dynamic transitions. These endeavours and others will surely accelerate a fruitful era of topological materials and the control of these materials for applications. ■

Young-Woo Son *is at the Korea Institute for Advanced Study, Seoul 02455, South Korea.*
e-mail: hand@kias.re.kr

1. Geim, A. K. & Kim, P. *Sci. Am.* **298**, 90–97 (2008).
2. Armitage, N. P., Mele, E. J. & Vishwanath, A. *Rev. Mod. Phys.* **90**, 015001 (2018).
3. Sie, E. J. *et al. Nature* **565**, 61–66 (2019).
4. Dirac, P. A. M. *Proc. R. Soc. Lond. A* **117**, 610–624 (1928).
5. Weyl, H. *Z. Phys.***56**, 330–352 (1929).
6. Herring, C. *Phys. Rev.* **52**, 365–373 (1937).
7. Soluyanov, A. A. *et al. Nature* **527**, 495–498 (2015).
8. Clarke, R., Marseglia, E. & Hughes, H. P. *Phil. Mag.* **38**, 121–126 (1978).
9. Kim, H.-J., Kang, S.-H., Hamada, I. & Son, Y.-W. *Phys. Rev. B* **95**, 180101(R) (2017).
10. Zeljkovic, I. *et al. Nature Nanotechnol.* **10**, 849–853 (2015).
11. Liu, Y. *et al. Nature Phys.* **10**, 294–299 (2014).

BIOCHEMISTRY

Signalling molecule reprograms metabolism

The signalling molecule nitric oxide protects the kidneys by reprogramming metabolism, and its levels are regulated by a two-component system in mice. These findings identify new targets for drug discovery. SEE LETTER P.96

CHARLES J. LOWENSTEIN

Acute kidney injury can lead to chronic renal failure, which causes fluid and electrolyte imbalances in the blood that require dialysis. Such injuries commonly involve ischaemia–reperfusion events, in which the blood supply to the kidney is temporarily restricted but then restored; this process generates toxic oxygen radicals that can cause renal inflammation and damage. Zhou *et al.*[1] report on page 96 that the signalling molecule nitric oxide[2,3] reprograms a metabolic pathway, and thereby limits ischaemic injury and protects renal function.

Nitric oxide is synthesized by a family of

© 2019 Springer Nature Limited. All rights reserved.

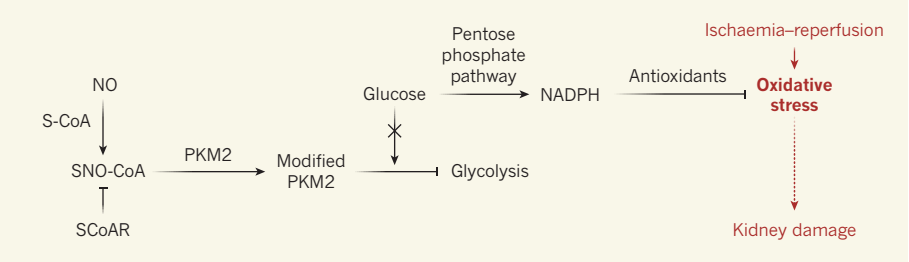


Figure 1 | Nitric oxide reprograms metabolism and limits oxidative stress. Zhou *et al.*[1] report that, in mice, the signalling molecule nitric oxide (NO) attaches to the molecule *S*-coenzyme A (S-CoA) to form *S*-nitroso-coenzyme A (SNO-CoA). This, in turn, delivers nitric oxide to the enzyme pyruvate kinase M2 (PKM2), modifying PKM2 and thereby inhibiting glycolysis — a metabolic pathway that consumes glucose. Glucose therefore enters another metabolic pathway, the pentose phosphate pathway, which generates NADPH, a cofactor used by antioxidants. The antioxidants can inhibit oxidative stress in kidney cells caused by a process called ischaemia–reperfusion, thus limiting damage to the kidneys (dotted arrow indicates damage limitation). The enzyme *S*-nitroso-coenzyme A reductase (SCoAR) removes nitric oxide from SNO-CoA, and so controls how much nitric oxide is available to modify target proteins.

enzymes called nitric oxide synthases (NOS), which fall into three groups: neuronal NOS, inducible NOS and endothelial NOS (eNOS). The molecule signals through several distinct mechanisms[4]. For example, it can interact with transition metals such as those in the haem group of guanylyl cyclase enzymes, which produce cyclic GMP — a messenger molecule involved in many biological processes. It can also combine with oxygen molecules to produce reactive nitrogen oxide species that, in turn, react with cysteine amino-acid residues on target proteins[5], forming modifications called *S*-nitrosothiols. Nitric oxide regulates a variety of physiological processes, including dilation of blood vessels (vasodilation), communication between neurons and the killing of disease-causing agents by the immune system.

Zhou and colleagues now show that nitric oxide protects kidneys from ischaemic damage. In particular, they observed that renal injury after ischaemia and reperfusion was worse in mice genetically engineered to lack eNOS than in wild-type mice. This result is consistent with previous findings that nitric oxide — not only nitric oxide produced in the body, but also that introduced from an external source[6] — can limit ischaemic injury in the kidneys[7], heart[8], brain[9] and other organs. The role of nitric oxide in these protective effects was not fully understood, but it has been proposed to act variously as an antioxidant[10], an anti-inflammatory agent[11] or a vasodilator[7].

The authors of the current study set out to identify the pathways by which nitric oxide protects against ischaemia. Using mass spectrometry, they discovered that one of the proteins most commonly modified by the molecule is pyruvate kinase M2, an enzyme that catalyses glycolysis (the metabolic pathway by which glucose is converted into energy). In a clever set of biochemical studies, they showed that nitric oxide modifies specific cysteine residues of pyruvate kinase M2. These modifications block the assembly of the active form of the enzyme, thereby inhibiting glycolysis. This is one of the key findings of the study: pyruvate kinase M2 is a target of nitric oxide.

Zhou *et al.* next genetically engineered mice so that their kidneys did not produce pyruvate kinase. The authors found that ischaemia causes less damage in these mice than in wild-type mice, consistent with the idea that pyruvate kinase mediates the protective effects of nitric oxide. But how?

The researchers used a technique called metabolic profiling to show that the kidney cells of mice lacking pyruvate kinase have high levels of products of the pentose phosphate pathway[12] — a metabolic pathway parallel to glycolysis that produces sugars called pentoses and the enzyme cofactor NADPH. NADPH acts in antioxidant systems to restore the function of proteins that have been damaged by oxidative stress in ischaemia[13]. The authors therefore conclude that nitric oxide inhibits pyruvate kinase and glycolysis, causing glucose levels to increase. The excess glucose spills over into the pentose phosphate pathway, generating high levels of NADPH, which shores up the antioxidant defences that limit renal injury (Fig. 1). This reprogramming of metabolism represents a major new aspect of nitric oxide biology.

How is nitric oxide conveyed to its renal-protein targets? Workers from the same group as Zhou and colleagues had previously identified[14] a two-component system that controls the availability of nitrosothiol groups in yeast. The first component is *S*-nitroso-coenzyme A, a molecule that donates nitric oxide groups to target proteins. The second component is an enzyme called *S*-nitroso-coenzyme A reductase, which removes nitric oxide from *S*-nitroso-coenzyme A. But does this binary system have any relevance to mammals?

To answer this question, Zhou *et al.* studied the impact of *S*-nitroso-coenzyme A reductase in mice during renal ischaemia and reperfusion. As expected, genetic deletion of the enzyme increased levels of *S*-nitrosylated proteins, protected mice from renal damage and prolonged survival compared with results in wild-type mice. Kidney levels of NADPH were also increased compared with levels of its oxidized form, $NADP^+$, as were levels of the antioxidant glutathione relative to its oxidized form, glutathione disulfide, confirming that protection occurs through the action of antioxidant defences. These exciting results show that *S*-nitroso-coenzyme A reductase acts *in vivo* in mammals to control nitric oxide signalling, which is the third major discovery of the study.

This work highlights important questions for further research. The authors' identification of a two-component system for regulating *S*-nitrosylation levels in renal injury raises the issue of what effect this system has on such regulation in normal physiological processes. How does this system function during other disorders, such as inflammation and cancer, which are also characterized by oxidant stress? And could pyruvate kinase M2 be a target for anti-ischaemic therapies?

Further work is also needed to identify how modification of pyruvate kinase M2 by nitric oxide protects cells — through inhibition of the enzyme's metabolic activity, or by inhibiting its other functions[15] (such as protein kinase activity and transcriptional co-activation)? Finally, Zhou *et al.* show that nitric oxide inhibits glycolysis in the setting of renal ischaemia, but it has previously been shown that it increases glycolysis in other settings[16]. Perhaps the activity of the newly discovered two-component regulatory system can explain previously puzzling aspects of nitric oxide biology, and might open up approaches for treating ischaemic injury in the kidney and other organs. ■

Charles J. Lowenstein *is in the Department of Medicine, Aab Cardiovascular Research Institute, University of Rochester Medical Center, Rochester, New York 14642, USA.*
e-mail: charles_lowenstein@urmc.rochester.edu

1. Zhou, H.-L. *et al. Nature* **565**, 96–100 (2019).
2. Nathan, C. *FASEB J.* **6**, 3051–3064 (1992).
3. Lowenstein, C. J. & Snyder, S. H. *Cell* **70**, 705–707 (1992).
4. Maron, B. A., Tang, S.-S. & Loscalzo, J. *Antioxid. Redox Signal.* **18**, 270–287 (2013).
5. Stamler, J. S. *et al. Proc. Natl Acad. Sci. USA* **89**, 444–448 (1992).
6. Johnson, G., Tsao, P. S. & Lefer, A. M. *Crit. Care Med.* **19**, 244–252 (1991).
7. Park, K. M. *et al. J. Biol. Chem.* **278**, 27256–27266 (2003).
8. Flögel, U., Decking, U. K. M., Gödecke, A. & Schrader, J. *J. Mol. Cell. Cardiol.* **31**, 827–836 (1999).
9. Endres, M. *et al. Proc. Natl Acad. Sci. USA* **95**, 8880–8885 (1998).
10. Erkens, R. *et al. Oxidat. Med. Cell. Longev.* **2018**, 8309698 (2018).
11. Liu, P., Yin, K., Nagele, R. & Wong, P. Y.-K. *J. Pharmacol. Exp. Ther.* **284**, 1139–1146 (1998).
12. Stincone, A. *et al. Biol. Rev. Camb. Phil. Soc.* **90**, 927–963 (2015).
13. Xiao, W., Wang, R.-S., Handy, D. E. & Loscalzo, J. *Antioxid. Redox Signal.* **28**, 251–272 (2018).
14. Anand, P. *et al. Proc. Natl Acad. Sci. USA* **111**, 18572–18577 (2014).
15. Israelsen, W. J. & Vander Heiden, M. G. *Semin. Cell Dev. Biol.* **43**, 43–51 (2015).
16. Almeida, A., Moncada, S. & Bolaños, J. P. *Nature Cell Biol.* **6**, 45–51 (2004).

This article was published online on 28 November 2018.

© 2019 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0770-2

Scalable energy-efficient magnetoelectric spin–orbit logic

Sasikanth Manipatruni[1]*, Dmitri E. Nikonov[1], Chia-Ching Lin[1], Tanay A. Gosavi[1], Huichu Liu[2], Bhagwati Prasad[3], Yen-Lin Huang[3,4], Everton Bonturim[3], Ramamoorthy Ramesh[3,4,5] & Ian A. Young[1]

Since the early 1980s, most electronics have relied on the use of complementary metal–oxide–semiconductor (CMOS) transistors. However, the principles of CMOS operation, involving a switchable semiconductor conductance controlled by an insulating gate, have remained largely unchanged, even as transistors are miniaturized to sizes of 10 nanometres. We investigated what dimensionally scalable logic technology beyond CMOS could provide improvements in efficiency and performance for von Neumann architectures and enable growth in emerging computing such as artifical intelligence. Such a computing technology needs to allow progressive miniaturization, reduce switching energy, improve device interconnection and provide a complete logic and memory family. Here we propose a scalable spintronic logic device that operates via spin–orbit transduction (the coupling of an electron's angular momentum with its linear momentum) combined with magnetoelectric switching. The device uses advanced quantum materials, especially correlated oxides and topological states of matter, for collective switching and detection. We describe progress in magnetoelectric switching and spin–orbit detection of state, and show that in comparison with CMOS technology our device has superior switching energy (by a factor of 10 to 30), lower switching voltage (by a factor of 5) and enhanced logic density (by a factor of 5). In addition, its non-volatility enables ultralow standby power, which is critical to modern computing. The properties of our device indicate that the proposed technology could enable the development of multi-generational computing.

Transistor technology scaling[1–3] has been enabled by controlling the conductivity of a semiconductor using an electric field applied across a high-quality insulating gate dielectric. This fundamental principle has remained largely unchanged since the seminal observations of Moore and Dennard et al.[4,5]. Yet in the past decade, transistor scaling has been enabled by direct improvements to the carrier transport[1,6,7], combined with superior electrostatic control[1–3,8]. In contrast to pure dimensional scaling[5], new transistor technologies have necessitated the use of strain[6], three-dimensional electrostatic gate control[2,8], manipulation of the effective carrier mass and band structure, and the gradual introduction of new materials for interface and work function control[9]. Despite the successful scaling in the size of transistors, voltage and frequency scaling have slowed[10]. Further decrease of voltage has been hampered by the Boltzmann limit of current control (60 mV for every change in current by a factor of 10 at room temperature). In response, a considerable effort to invent, demonstrate and benchmark beyond-CMOS devices got underway[11–13]. This effort includes alternative computing devices based on electron spin, electron tunnelling, ferroelectrics, strain and phase change[12,13] (see Methods for beyond-CMOS logic device options). However, a technologically suitable computational logic device that has superior energy efficiency, high logic density (that is, computed functions per unit area), non-volatility (to counteract leakage power) and efficient interconnects has remained elusive. The importance of these considerations has become evident during extensive modelling, benchmarking and evaluation of more than 25 beyond-CMOS device proposals[12,13]. With these considerations in view, we propose and demonstrate the building blocks for a new logic device that enables (1) voltage scaling, (2) scalable interconnects, (3) energy scaling and (4) the potential for multi-generational dimensional scaling.

Beyond-CMOS devices for replacing or enhancing the electronic transistor

Collective state switching devices are potential candidates for replacing or enhancing transistors. A collective state switch operates by the reversal of the material's order parameter (such as ferromagnetism, ferroelectricity and ferrotorodicity)[13] from Θ to $-\Theta$. It addresses sub-10-nm miniaturization by using collective order parameter dynamics, overcoming the 'Boltzmann tyranny', which is inherent to conductivity modulation, and providing a non-volatile nature to the computer. It is well documented that the 'Boltzmann tyranny' and leakage are the central challenges in traditional CMOS devices[1,2]. Logic based on collective state switching devices is a leading option for computational advances beyond the modern CMOS era owing to its (1) potential for superior energy per operation, (2) higher computational logical density and efficiency (that is, fewer devices required per combinatorial logic function) owing to the use of majority gates[14], (3) non-volatile memory-in-logic and logic-in-memory capability[15] and (4) amenability to traditional and emerging architectures (for example, neuromorphic[16] and stochastic computing[17]).

Among these possible collective state order parameters, ferroelectricity and multiferroicity are the preferred collective states for computing[13] owing to (1) the presence of a controllable, localized and phenomenologically strong carrier, the spontaneous dipole; (2) the switching efficiency of a ferroelectric with respect to the stability of the switch is given by the energy barrier per unit volume, $\lambda = E_{sw}/\Delta E(\Theta)$, where $\Delta E(\Theta)$ is the energy barrier relative to the stable state and E_{sw} is the total energy dissipated in switching; lower values of λ enable computing switches to operate at lower energies for a given energy barrier.

A vital consideration for a new technology is the need for highly compact nanoscale interconnects. While ferroelectric switching and the

[1]Components Research, Intel Corporation, Hillsboro, OR, USA. [2]Intel Labs, Intel Corp., Santa Clara, CA, USA. [3]Department of Materials Science and Engineering, University of California, Berkeley, Berkeley, CA, USA. [4]Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [5]Department of Physics, University of California, Berkeley, Berkeley, CA, USA. *e-mail: sasikanth.manipatruni@intel.com

© 2019 Springer Nature Limited. All rights reserved.

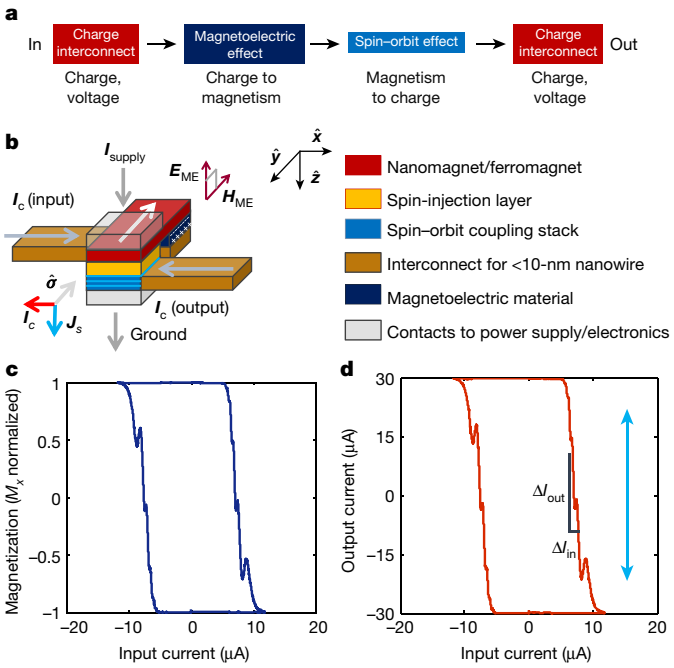


Fig. 1 | MESO logic transduction and device operation. a, Transduction of state variables for a cascadable charge-input and charge-output logic device. The magnetoelectric effect transduces the input information to magnetism, and the spin–orbit effect in a topological material transduces the magnetic state variable back to charge. **b**, MESO device formed with a magnetoelectric capacitor and a topological material. The device comprises a spin-injection layer for spin injection from the ferromagnet to the topological material, an interconnect made of a conductive material, and contacts to the power supply and ground. The logical state of the charge input (current in the $+\boldsymbol{x}$ direction) is inverted by the operation shown to charge output (current in the $-\boldsymbol{x}$ direction). Power for energy gain is injected from the power supply (arrows). Transduction mechanisms are calculated with magnetoelectric-vector SPICE models (see Methods and Supplementary Information). The white arrow represents the magnetization direction of the ferromagnet. Grey arrows represent electric currents at the input and output, power supply and ground. Injection of the power supply current allows for energy gain, large signal gain and the ability to drive larger output devices. **c**, Magnetoelectric transfer function, showing conversion of the charge input to ferromagnetic magnetization. **d**, Spin–orbit transfer function, showing conversion of a state to charge output. The response of the device is indicated for small signal gain (black line) and the full signal range (−15 μA to 15 μA; blue arrow). See Supplementary Fig. 1 for the two operating states of the MESO inverter device.

accompanying magnetoelectric switching of ferromagnets are perhaps the most energy-efficient charge-driven switching phenomena at the nanoscale and at room temperature, an efficient way to read out the state has been lacking. The discovery of strong spin–charge coupling in topological matter via a Rashba–Edelstein or topological two-dimensional electron gas[18–25] enables this proposal for a charge-driven, scalable logic computing device.

Spin–orbit logic device with magnetoelectric input signal nodes

We propose a logic computing device with magnetoelectric switching nodes and spin–orbit-effect readout operating at 100 mV, with an electrical interconnect. The magnetoelectric spin–orbit (MESO) device comprises two technologically scalable transduction mechanisms: ferroelectric/magnetoelectric switching[26–30] and topological conversion of spin to charge[19–24]. The device interfaces with electrical interconnects and is therefore charge-/voltage-driven and produces a charge/voltage output (Fig. 1a). The MESO device (Fig. 1b) comprises a magnetoelectric switching capacitor, a ferromagnet and a spin-to-charge conversion module (see 'Material requirements for 1–10-aJ-class MESO logic'). In Fig. 1b, when the input interconnect carries a positive current (current flowing in the $+\boldsymbol{x}$ direction), an electric field is set up in the magnetoelectric capacitor in the $-\boldsymbol{z}$ direction (into the plane). The resulting magnetoelectricity (represented as an effective field $\boldsymbol{H}_{ME}$), which may be comprised of an electrically controlled exchange bias or exchange anisotropy, switches the nanomagnet to the $-\boldsymbol{y}$ direction[27–30] (see Supplementary Information section A for details).

The readout (detection) of the state of the switch is enabled by the ongoing advances in spin-to-charge conversion using topological or high-spin–orbit-coupling (SOC) materials. A supply current is injected into the device, causing a flow of spin-polarized electrons from the ferromagnet into the SOC material. Owing to SOC spin-to-charge transduction (Fig. 1b), a charge current is generated at the output, in this case in the $-\boldsymbol{x}$ direction. Hence, the input charge state (positive voltage and current) is inverted by the MESO logic gate at the output.

We applied spin/magnetoelectric circuit theory[31,32] (see Supplementary Information section B) combined with stochastic magnetization dynamics solvers[32] (see Supplementary Information sections C and D) to obtain the transfer characteristics of the MESO logic device. We further used rigorous integrated-circuit solvers to validate and benchmark against spin-logic examples (Supplementary Information sections E and F). SPICE (simulation program with integrated circuit emphasis) circuit solvers were developed to incorporate the effects of: (1) magnetoelectric switching, (2) all the energy sources and dissipation elements (Supplementary Information sections G and H), (3) Landau–Khalatnikov dynamics of ferroelectric switching (Supplementary Information section I) and (4) peripheral charge circuitry (Supplementary Information section J). Details on simulations of energy scaling to <10 aJ (1 aJ = 10^{-18} J) using the power boundary method and component-level energy calculations are presented in Supplementary Information section K. The MESO inverter transfer functions, with magnetic and electric hysteresis, are shown in Fig. 1c, d. The input current I_{in} of the magnetization transfer function (Fig. 1c) relates the magnetoelectric stimulus with magnetization switching. It shows a small-signal gain ($d\hat{\boldsymbol{m}}/dI_{in}$, where $\hat{\boldsymbol{m}}$ is the magnetic moment unit vector for the nanomagnet's magnetization), which is advantageous for noise rejection. A large-signal gain of the output current, I_{out} (ratio I_{out}/I_{in}) is generated and controlled by the supply current (I_{supply}). A small-signal gain (dI_{out}/dI_{in}) of the device during switching can be seen in Fig. 1d. We show a scheme of the proposed short-/long-range interconnect in Fig. 2a, where the charge output of one MESO stage drives a charge current to switch the input of the next MESO stage[32,33]. Bidirectional logic switching of a cascaded six-stage MESO inverter chain is described in Supplementary Information section L.

Transduction mechanisms for the MESO device

We identified a scalable way to transduce the spin state of a nanomagnet to a charge state via spin–orbit effects[19–26,33–36], such as the interface Rashba–Edelstein effect (IREE) and spin-momentum locking in topological insulators. It has recently been shown that spin currents can be converted to charge currents that preserve the information encoded in spin polarization[34–37] (using resonant spin pumping in the quasi-static non-local spin-valve configuration[33]). Figure 2b shows how a current through a nanomagnet produces injection of spin-polarized electrons into a stack composed of materials with a high SOC coefficient (for example, Bi/Ag[22,37], topological insulators[21,24,34,35], oxides and two-dimensional materials[25,36]). In Fig. 2b, when $\hat{\boldsymbol{m}}$ is pointing in the $\hat{\boldsymbol{y}}$ direction and the flow of the injected spin current is $\boldsymbol{J}_s = J_s\hat{\boldsymbol{z}}$, with injected spin polarization along the $+y$ direction, charge current $\boldsymbol{I}_c$ is generated in the $\hat{x}$ direction. When the nanomagnet reverses to the $-\hat{\boldsymbol{y}}$ direction and the flow of injected spin current is still $\boldsymbol{J}_s = J_s\hat{\boldsymbol{z}}$, but with injected spin polarization along the $-y$ direction, a charge current $\boldsymbol{I}_c$ is generated in the $-\hat{x}$ direction. Hence, the magnetization direction of the nanomagnet is transduced into the direction of the electric current.

© 2019 Springer Nature Limited. All rights reserved.

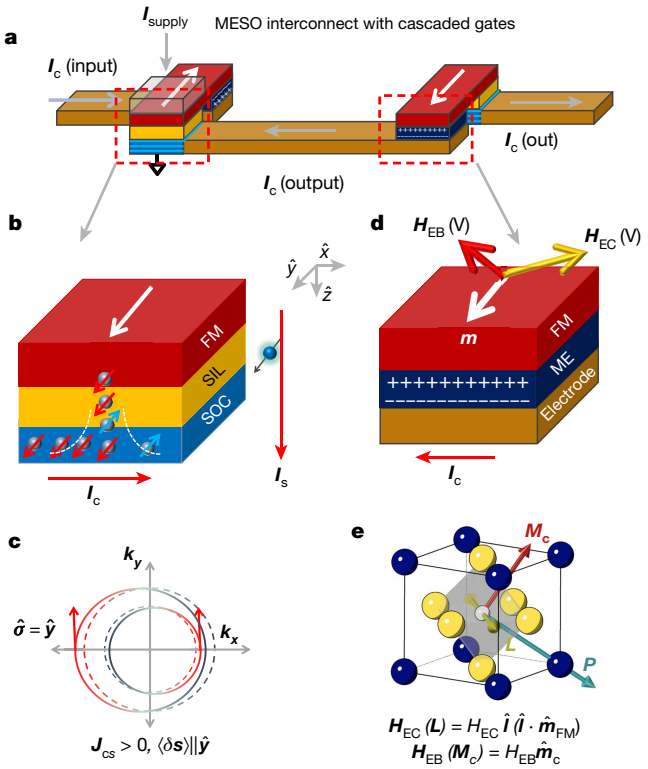


Fig. 2 | Operating mechanisms for MESO logic. a, A low-voltage-charge-based MESO interconnect with cascaded logic gates. Two inverters are chained together to form an interconnect. Arrows show the directions of the input and output currents of the device. Materials are as in Fig. 1. **b**, Operating mechanism for spin-to-charge conversion using a high-SOC material (SOC). A spin injection layer (SIL) is used where needed by the materials' interfaces. Spins injected from the ferromagnet (FM) in the $+\boldsymbol{z}$ direction with spin polarization along the $+\boldsymbol{y}$ (in-plane) direction cause a topologically generated charge current in the SOC layer. Small red and blue arrows indicate up and down spins, respectively, injected from the magnet. The large red arrows show the directions of the charge ($\boldsymbol{I}_c$) and injected spin ($\boldsymbol{I}_s$) currents. **c**, Schematic of the $\boldsymbol{k}$-space for spin-to-charge conversion at a two-dimensional electron gas with high SOC. Injecting a spin current polarized along the $+\boldsymbol{y}$ direction overpopulates the Fermi surface on one side of the topological material compared to the other side. This generates a net charge current in the $\boldsymbol{x}$ direction. The conversion has the right symmetry to convert the information of the ferromagnet to the charge current output. The dashed and solid lines depict the Fermi surface of the material before and after spin injection, respectively. Injected spin density $\langle \delta \boldsymbol{s} \rangle$ along the $+y$ direction leads to charge current $\boldsymbol{J}_{cs} > 0$. **d**, Operating mechanism for a magnetoelectric (ME) material. A ferromagnet is coupled via exchange/strain to the magnetoelectric material. $\boldsymbol{H}_{EB}$ and $\boldsymbol{H}_{EC}$ are the exchange bias and the exchange coupling from the magnetoelectric material to the ferromagnet, respectively, and $\boldsymbol{m}$ is the magnetization of the ferromagnet. **e**, A classic multiferroic–magnetoelectric material, $BiFeO_3$, is shown with the order parameters: polarization ($\boldsymbol{P}$), antiferromagnetism ($\boldsymbol{L}$) and weak canted magnetization ($\boldsymbol{M}_c$). The electric-field setup in a generic magnetoelectric/ferroelectric material produces exchange bias, coupling and anisotropy modulation for magnetostrictive effects. Yellow and blue spheres depict Bi and Fe atoms in a canonic room-temperature multiferroic $BiFeO_3$. $\hat{\boldsymbol{m}}_{FM}$, $\hat{\boldsymbol{m}}_c$ and $\hat{\boldsymbol{l}}$ represent the unit vectors of magnetization of the coupled layer, the magnetization of the canted spins and the antiferromagnetic axis, respectively. In general, the magneto-electric field generates an exchange coupling along the axial direction of $\hat{\boldsymbol{l}}$, the AFM axial direction and exchange bias, the direction of weak ferromagnetism $\hat{\boldsymbol{m}}_c$.

The spin–orbit mechanism responsible for spin-to-charge conversion at the interface is described by the Hamiltonian

$$H_R = \alpha_R (\boldsymbol{k} \times \hat{\boldsymbol{z}}) \cdot \hat{\boldsymbol{\sigma}} \quad (1)$$

where $\alpha_R = (k_{F+} - k_{F-})\hbar^2/2m$ is the Rashba coefficient ($\hbar$ is the Planck constant), $\boldsymbol{k}_{F+}$ and $\boldsymbol{k}_{F-}$ are the Fermi vectors of the two spin-split bands, $\hat{\boldsymbol{z}}$ is the unit vector normal to the interface, $\hat{\boldsymbol{\sigma}}$ is the vector of the Pauli spin matrices and $\boldsymbol{k}$ is the momentum of the electrons. In a simple model based on two Fermi contours in the Rashba electron gas (Fig. 2c), the density of spin polarization along the y axis, $\delta s_{y\pm}$ (Fig. 2b, c), and the charge current density along the x axis, $j_{cx\pm}$, can be related as[20,23]

$$\delta s_{y\pm} = \pm \frac{m}{2e\hbar k_{F\pm}} j_{cx\pm} \quad (2)$$

which yields the relation between spin density (per unit area) and charge current (per unit width) in a two-dimensional Rashba electron gas:

$$j_{cx} = \frac{e\alpha_R}{\hbar}\langle \delta s \rangle_y = \frac{\alpha_R \tau_s}{\hbar} j_{sy} = \lambda_{IREE} j_{sy} \quad (3)$$

where the relation between spin current and spin polarization is determined by the spin relaxation time τ_s as $j_s = e\delta s/\tau_s$, where e is the electron charge. For a pure helical ground state in topological systems, $\lambda_{IREE} = V_F\tau$, where V_F is the Fermi velocity and τ is the relaxation time for the spin distribution at an out-of-equilibrium interface. This results in the generation of a charge current in the interconnect that is proportional to the spin current (Fig. 2c). The transduction relates the linear charge current density j_{cx} (in units of ampere per metre) and the areal spin current density j_{sy} (spin current flowing along the z direction, comprised of spins oriented along the y direction; in units of ampere per square metre); see Supplementary Information section M and Supplementary Fig. 16.

Magnetoelectricity provides a highly energy-efficient mechanism for logic switching with intrinsic switching energy given by

$$E_{ME} = 2P_s V_c \quad (4)$$

where P_s is the switched polarization and V_c the critical voltage for switching. To the best of our knowledge, magnetoelectric/ferroelectric switching is the most energy-efficient mechanism at room temperature that scales to lateral dimensions of 10 nm and retains a stable collective order parameter. The switching mechanism for magnetoelectric switching of a ferromagnet is shown Fig. 2d, and a canonical room-temperature multiferroic magnetoelectric ($BiFeO_3$) is illustrated in Fig. 2e. In general, magnetoelectric switching can be accomplished by coupling the ferroelectricity/ferroelasticity to antiferromagnetism and/or a weak canted magnetic moment. The intrinsic switching energy for ferroelectric/magnetoelectric switching can approach 1 aJ per bit (about 30 times lower than the switching energy of advanced CMOS devices) by scaling the switched polarization to about 10 $\mu C\,cm^{-2}$ and switching voltages to 100 mV (Please see Supplementary Fig. 24 for low-voltage ferroelectric characterization of SRO/20 nm LBFO/SRO heterostructure). Both of these metrics are within the reach of experimental room-temperature materials (as shown in Table 1).

Miniaturization and scaling laws for MESO logic

We now derive and apply the scaling laws for magnetoelectric and spin–orbit transductions. For spin-to-charge conversion using inverse SOC (ISOC), the efficiency improves with reducing the width of the magnet, a highly desirable scaling feature. In the presence of topological coupling between spin and charge states in a Rashba system, we can write

$$\boldsymbol{I}_c = \frac{1}{w}\lambda'_{ISOC}(\hat{\boldsymbol{\sigma}} \times \boldsymbol{I}_s) \quad (5)$$

where w is the width of the magnet (the minimum feature size for the device), where λ'_{ISOC} is the effective SOC conversion length. The recent discovery of two-dimensional high-SOC systems indicates that the effective λ_{IREE} can be as high as 6 nm in Bi_2Se_3[34,35] and $LaAlO_3$/$SrTiO_3$[25,36,38]. To assess the suitability of spin–orbit logic for progressive

© 2019 Springer Nature Limited. All rights reserved.

Table 1 | Device and material targets to enable 1–10-aJ-class MESO logic

	Device figure of merit	Nominal target	Material figures of merit	Nominal target (for 1–10 aJ per switch)
SOC materials	Spin-to-charge conversion, I_c/I_s	>50%	λ_{IREE}	>5 nm[25,35,36,38]
	Source resistance	>10 kΩ	Resistivity	>10 mΩ cm (refs [25,26,35,36,38])
Magnetoelectrics	Dimensions	$<10 \times 10 \times 10$ nm^3	Charge/area (for 1–10 aJ target)	0.5–5 μC cm^{-2} (ref. [42])
	Equivalent capacitance/area	<100 fF μm^{-2}	Coercive field	100–500 kV cm^{-1} (refs[27,30])
	Switching voltage	0.1 – 0.3 V	Magnetoelectric coefficient, α_{ME}	10 C^{-1} (refs [27,28])
	Write error rate	$<10^{-12}$	Reliability	$>10^{15}$
Interconnect	Resistance/ length	0.1–5 kΩ μm^{-1}	Resistivity	4–200 μΩ cm at 10 nm width[3]
	Capacitance/ length	10–100 aF μm^{-1}	Interlayer dielectric (dielectric constant)	1–10 (ref. [3])
	Peak currents	10–100 μA per magnet	Electromigration limit	>25 MA cm^{-1} at 10 nm width
Nanomagnet	Size	20 nm × 30 nm		
	Magnetic stability (barrier, Δ/kT)	40	Magnetization, M_s	<500 MA cm^{-1}
	Spin injection	>80%	Spin polarization	>80%

miniaturization (Moore's law), we show that the energy required to switch the device decreases with dimensional scaling of the device. This can be attributed to an improvement in the spin-to-charge conversion efficiency, η_{SOC}, with a reduction in the width W_{FM} of the nanomagnet ($\eta_{SOC} \propto 1/W_{FM}$) and a reduction in the switched charge of the magnetoelectric/ferroelectric node, Q, with areal scaling ($Q \propto W_{FM}^2$) . The energy required to switch a single MESO logic unit is given by

$$\begin{aligned} E_{MESO} &= E_{CME} + E_{IC} + E_{ISOC} + E_{RT} + E_{SG} \\ &= C_{me}V_{me}^2\left(1 + \alpha\frac{W_{FM}}{\lambda'_{SOC}}\right) \end{aligned} \quad (6)$$

C_{me} is the equivalent capacitance, V_{me} is the switching voltage and α is the conversion factor. (see Supplementary Information sections G–K for detailed analytical and numerical energy calculations of the intrinsic magnetoelectric energy, E_{CME}, interconnect losses, E_{IC}, and losses in the spin-to-charge conversion layer, E_{ISOC}, in the driving electronics, E_{RT}, and in the supply–ground path, E_{SG}). Figure 3a shows the strong, cubic scaling of the MESO energy (E_{MESO}), where the energy is reduced by a factor of 8 for every reduction by a factor of 2 in feature size. The excellent scalability of MESO logic allows the switching energy of the MESO logic to approach 1 aJ per bit. Magnetoelectric switching also allows strong voltage scaling in energy per bit, where progressive voltage reduction enables lowering of the switching energy. Figure 3c shows a combination of scaling in energy/switching via voltage and effective IREE length towards the 0.1–10 aJ range.

Low-voltage (100 mV) charge interconnects for scaling below 10 nm

We now address one of the most demanding aspects of new computing technology: the interconnects that connect the devices at the nanoscale[39,40]. MESO logic can address the interconnect scaling problem, which has emerged as a major limitation when the width of the electrical wires reached <20 nm. Experimental data for highly scaled interconnects show that the resistivity of electrical wires increases according to the Mayadas–Shatzkes scaling law[41]

$$\rho = \rho_0\left[1 + \frac{3\lambda_{ebulk}}{8t}\left(1 + \frac{p}{2}\right) + \frac{3\lambda_{ebulk}}{2D}\left(\frac{r}{1-r}\right)\right] \quad (7)$$

(where ρ_0 is the bulk resistivity, λ_{ebulk} is the electron mean free path, p is the specularity, r is the reflection parameter from grain boundaries, t is the thickness of the film and D the grain size) as the critical interconnect dimensions approach the electron mean free path. A second scaling issue with electrical interconnects is the high capacitance per unit length. Hence, it is of great interest to demonstrate a logic technology compatible with high-resistivity and high-capacitance interconnects.

MESO logic enables the development of a low-voltage charge interconnect that is amenable to highly scaled integrated circuits that comprise wires of 10–30 nm width. We show that MESO logic can tolerate the use of nanometallic interconnects with high resistivity (>1 mΩ cm; a 20–100-times less stringent requirement for the conductance of small-width interconnect material) and capacitance (>10 fF μm^{-1}; a 100-times less stringent requirement for the capacitance of the interconnects); see Supplementary Information section N. Figure 3b shows the dependence of switching time on the interconnect length at a metal resistivity of 100 μΩ cm and an effective 30 nm × 30 nm cross-sectional area of 900 nm^2. The switching speed of the MESO device scales linearly with interconnect length up to 1,000 nm at a line resistance of 100 Ω μm^{-1} and a capacitance of 100 aF μm^{-1} (see Supplementary Information section N). This would represent a substantial relaxation in the requirements placed on nanometallic interconnects compared to CMOS devices. This is in contrast to spin interconnects, where the signal, and thus the switching speed, degrade as $\exp(-x/L_{sf})$, where L_{sf} is the spin-flip length and x is the direction of transport. Hence, MESO logic alleviates the traditional problem of interconnects used for spin logic and allows continued scaling of metallic and semiconducting wires.

When compared with the leading beyond-CMOS and highly scaled advanced CMOS technologies, the MESO device provides sizeable gains in areal logic density, energy of operation and computational throughput. In Fig. 4a, we compare the energy delay and the logic density (area per function) of MESO logic with leading beyond-CMOS transistor technologies and highly scaled advanced CMOS devices. The majority logic gate, a universal gate that (together with NOT) can implement all Boolean logic functions, is used to build complex spin-logic functions, such as a 32-bit adder or a 32-bit arithmetic logic unit (ALU)[12], with the results shown in Fig. 4b (also see Supplementary Information section F). The proposed MESO logic device enables competitive energy delay performance compared to leading beyond-CMOS devices, while also allowing non-volatility. The MESO logic device enables considerable improvement compared to other spintronic devices owing to magnetoelectric switching. Gains in the delay are due to the use of electrical (rather than spin-based) interconnects and very compact majority-gate circuits (see Fig. 4c). The MESO device also enables energy reduction compared to CMOS logic operating at very low (0.3 V) supply voltages, owing to its ability to switch at even lower supply voltages (0.1 V). The speed of MESO logic units is comparable to that of low-power, low-leakage CMOS devices (0.3 V supply voltage). We note that at a low logic activity factor and intermittent usage, the non-volatility of the MESO device can offer further advantages compared to CMOS devices by eliminating standby power dissipation and enabling instant operation from standby. Figure 4c shows the advantage of the MESO device in terms of areal logic density compared to advanced CMOS technology.

© 2019 Springer Nature Limited. All rights reserved.

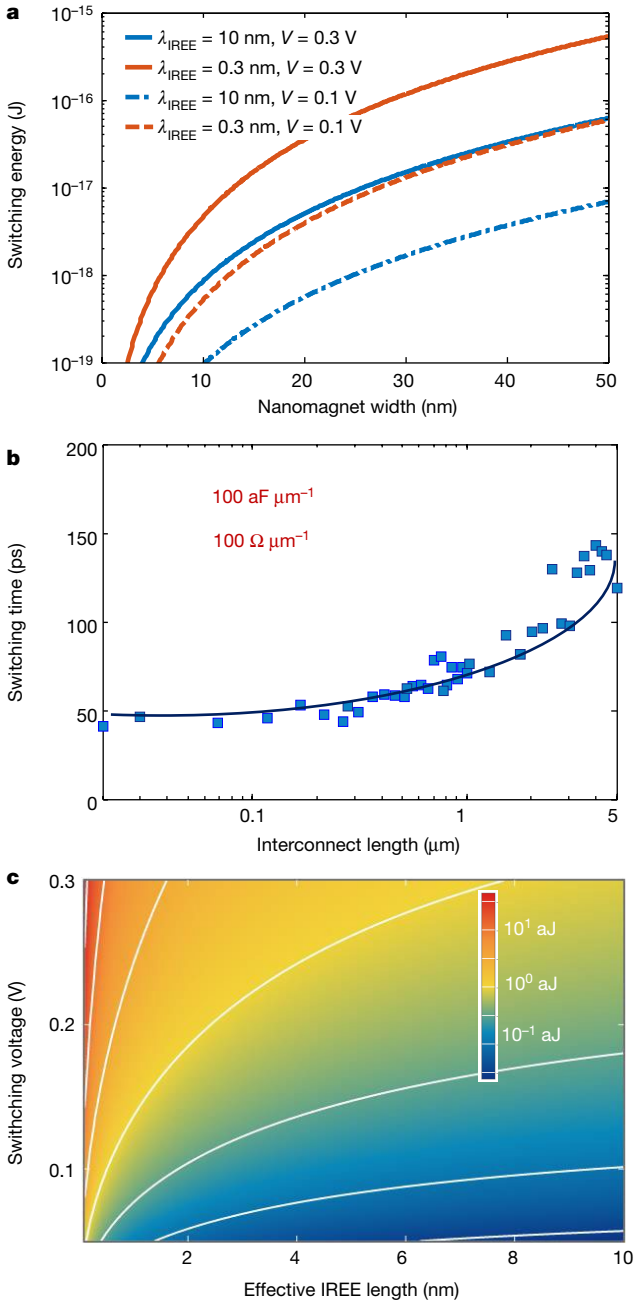


Fig. 3 | Energy and delay of the MESO device. Dimensional scaling of MESO logic shows improvement in switching energy as the devices get smaller. **a**, Cubic scaling with magnet width. **b**, Effect of the length of the interconnect (30 nm × 30 nm cross-section) on the (inverter) device performance, obtained from interconnect modelling of the MESO device. See Supplementary Information for the effects of the capacitance and length on device performance. We simulated the interconnect with 100 Ω μm^{-1} resistance and 100 aF μm^{-1} capacitance per unit length. **c**, Intrinsic switching energy of MESO (colour bar) with improvements in switching voltage and SOC strength (effective IREE length).

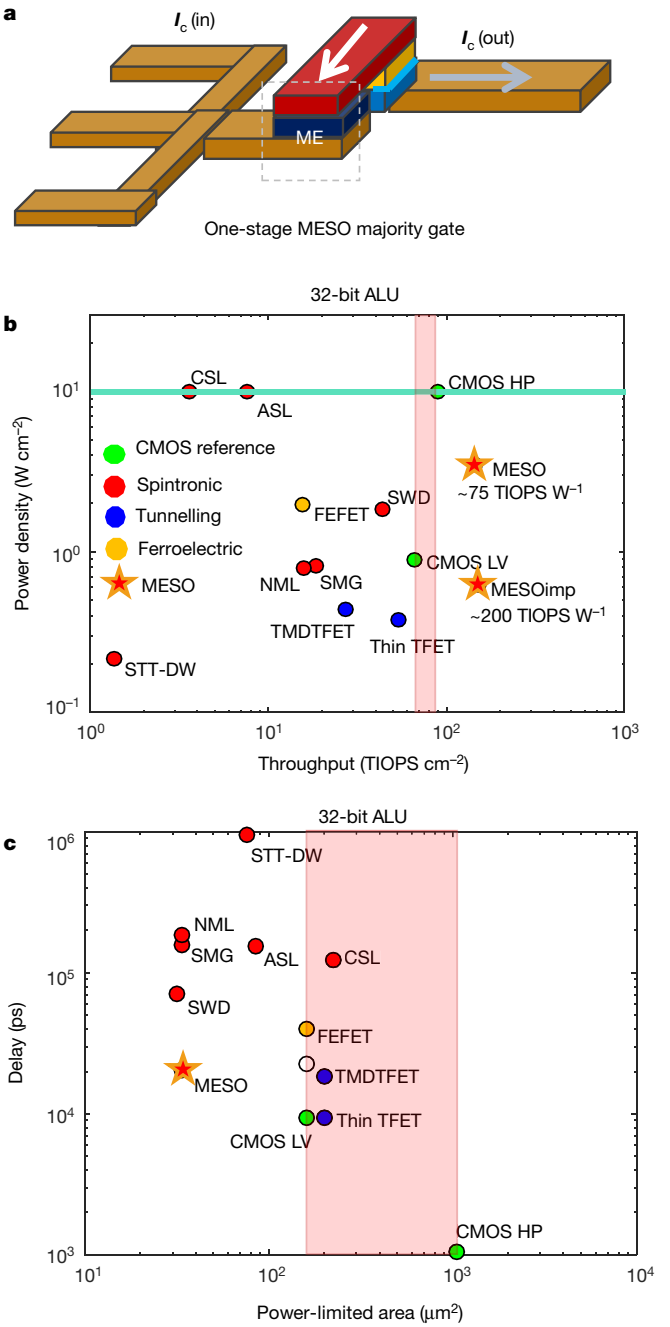


Fig. 4 | Performance and area of MESO device in comparison with advanced CMOS and leading beyond-CMOS devices. a, Scheme of a MESO majority logic gate used for benchmarking. **b**, Power per unit area versus throughput (that is, number of 32-bit ALU operations per unit time and unit area, in units of tera-integer operations per second; TIOPS) for CMOS and beyond-CMOS devices. The constraint of a power density not higher than 10 W cm^{-2} is implemented, when necessary, by inserting an empty area into the optimally laid out circuits. **c**, Delay of 32-bit ALU operation versus the power-limited area for CMOS and beyond-CMOS devices. Power-limited area comprehends the increase in area of the computational logic to meet the power density contraint of 10 W cm^{-2} in **b**. STT-DW, spin-transfer-torque domain-wall device; ASL, all-spin-logic device; CSL, charge spin logic; NML, nanomagnetic logic; SMG, spin majority gate; SWD, spin wave device; CMOS HP, high-performance CMOS at 0.73 V supply; CMOS LV, low-power CMOS operating at 0.3 V supply; FEFET, ferroelectric FET; Thin TFET, 2D-material vertical tunnel FET; TMDTFET, transition-metal dichalcogenide tunnel FET; see Methods.

Experimental progress on magneto–electrics and spin–orbit transduction

We now turn to the initial experimental manifestations of the two central concepts of the MESO device and conclude with a summary of innovations required to meet the MESO target of 100 mV and 1 aJ per bit. First we demonstrate a local-spin-injection device in which spin-polarized charge current I_s is injected from the ferromagnet (CoFe) into a spin–orbit material (Pt) (see Supplementary Information section P for the fabrication method and the device cross-section).

An open-circuit charge voltage V_{oc} that depends on the spin polarization of the injected spin current is measured across the Pt wire. The equivalent spin-to-charge-conversion resistance ($R_{SCC} = V_{oc}/I_s$) is plotted in

© 2019 Springer Nature Limited. All rights reserved.

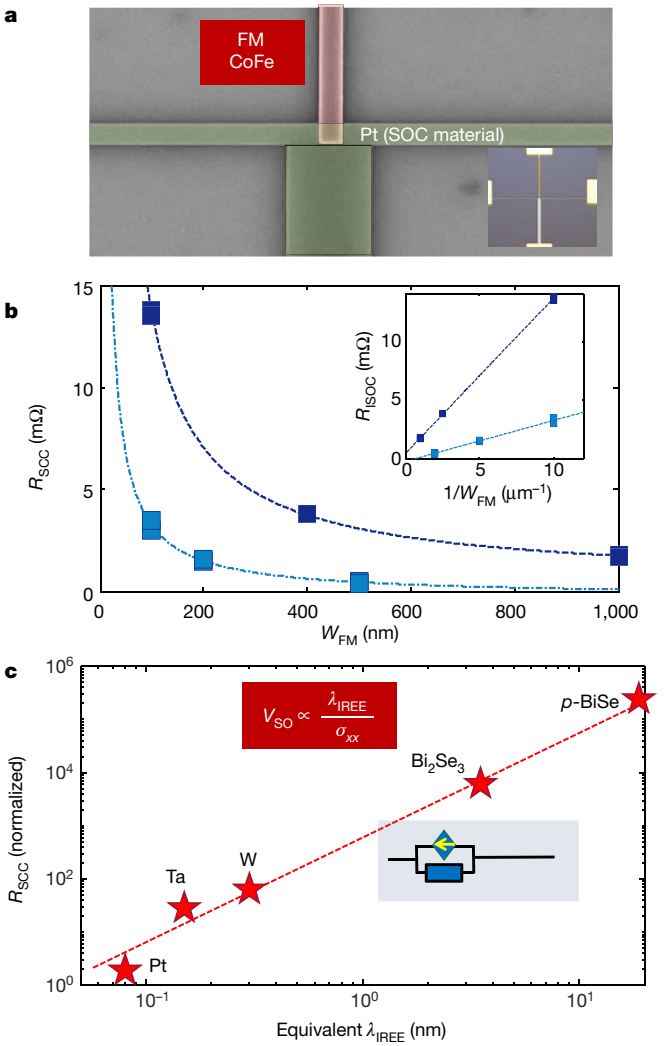


Fig. 5 | Spin–orbit readout for the MESO device. a, A proof-of-concept device for a spin–orbit readout mechanism in which a spin in a polarized charge current is passed into a spin–orbit material (Pt) from a CoFe ferromagnet. The inset shows a microscope image of the device. **b**, Charge readout of the spin state of the ferromagnet, measured using a non-local resistance (lateral voltage divided by vertical current). Shown are measured data from devices with various CoFe width and fixed Pt width of 100 nm (dark blue) and 400 nm (light blue). The inset shows linear fits to $1/W_{FM}$. **c**, Projected improvement in the spin–orbit readout through the use of topological insulators with low conductivity and high spin–orbit effects. V_{so} and σ_{xx} are the generated voltage and conductivity, respectively.

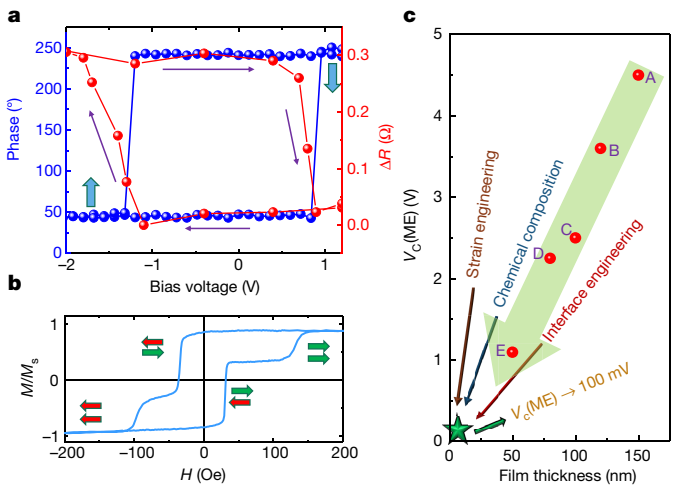


Fig. 6 | Progress of magnetoelectric transduction via MESO towards a voltage of 100 mV. a, Piezoelectric loop (blue) and applied-voltage-dependent resistance modulation (ΔR; red) of a Pt/CoFe/Cu/CoFe/LBFO/LSMO magnetoelectric spin-valve device depict 1 V magnetoelectric switching. The vertical axis shows the phase of the ferro-electric polarization response signal with respect to the input voltage. Purple arrows indicate the sweep direction in the measurement. **b**, Magnetization M_s of a giant magnetoresistive stack, normalized with the magnetization M of a thin film, as a function of magnetic field strength H. The bottom CoFe electrode of the giant magnetoresistive stack is exchange-coupled to $LaBiFeO_3$, showing exchange-induced anisotropy enhancement and exchange bias. **c**, Voltage scaling of magnetoelectric switching as a function of multiferroic film thickness. Data points A–D correspond to $BiFeO_3$ films of varying thickness with a $SrRuO_3$ bottom electrode. The switching voltage of the MESO device V_c(ME) is lowered down to 1.0 V using partial chemical substitution of Bi^{3+} (with La^{3+}), interface engineering (electrodes formed with $La_{0.7}Sr_{0.3}MnO_3$) and thickness scaling down to 40 nm. The ultimate goal of the MESO project is to be able to switch magnetoelectrically at 100 mV.

Fig. 5b. To study the dimensional scaling of the device, we measured R_{SCC} with two sets of test devices. Figure 5b shows measured data from devices with various CoFe widths W_{FM}. Test chip A (dark-blue line with Pt width 100 nm) comprises five different device geometries with dimensions $W_{FM} = 100$ nm, 100 nm, 400 nm, 1,000 nm and 1,000 nm and test chip B (light-blue line with Pt width is 400 nm) has five different device geometries with dimensions $W_{FM} = 100$ nm, 100 nm, 200 nm, 200 nm, 500 nm and 500 nm. We observe a dimensional scaling law, in accordance with the above expression for SOC spin-to-charge conversion; see inset for a linear fit to $1/W_{FM}$. The equivalent spin-to-charge conversion resistance scales favourably for high-resistivity topological materials as $R_{SCC} \propto \lambda_{IREE}/\rho$. For high-resistivity spin–orbit systems (see Table 1) we expect that R_{SCC} can be enhanced considerably. Combined with the areal scaling of the magnetoelectric capacitor, this provides a total device energy scaling of W_{FM}^3.

To illustrate the past year's remarkable progress in magnetoelectric transduction, we present in Fig. 6a–c representative magnetoelectric switching data of a CoFe/Cu/CoFe spin valve that is exchange-coupled to a 45-nm-thick $La_{0.1}Bi_{0.9}FeO_3$ (LBFO) layer using an applied voltage of only 1–2 V—although quasi-static ferroelectric switching has been achieved down to about 150 mV using LBFO as the magnetoelectric layer (see Supplementary Information section Q for details). Figure 6a shows the change in the resistance of the CoFe/Cu/CoFe spin valve as a function of voltage applied across the multiferroic LBFO thin film, along with the piezoelectric loop of the (CoFe/Cu/CoFe)/LBFO/LSMO capacitor structure. It is clear that both ferroelectric and concurrent magnetoelectric switching occur at about 1.0 V. This is enabled by the strong exchange coupling of the bottom CoFe layer of the CoFe/Cu/CoFe spin valve to the canted antiferromagnetic LBFO surface, which is evident from the magnetic hysteresis loop in Fig. 6b. Such an exchange-bias coupling can be reversed by an out-of-plane electric field. Figure 6c shows how rapidly this switching voltage has been decreasing over the past year, primarily through materials engineering of the switching behaviour of the $BiFeO_3$ layer, as well as through systematic thickness reductions. The ferroelectric saturation polarization and the switching voltage of $BiFeO_3$ can be tuned by the doping level of the rare-earth element, such as La or Sm[42]. The potential for further reductions in the switching voltage is illustrated through the quasi-static piezoelectric switching loop of a 20-nm-thick LBFO layer in contact with symmetric $SrRuO_3$ top and bottom electrodes, demonstrating a switching voltage of 130–150 mV (see details in Supplementary Information section Q). Further reductions in the switching voltage, which is the immediate focus of our research, should be possible through reduction of the LBFO film thickness and further careful tuning of the composition such that the polar distortion is delicately tuned to be as low as possible. Independent work by our group and others has shown that both the ferroelectric and antiferromagnetic orders are stable down to at least a few nanometres (see Supplementary Information). A proof of concept for an ultralow-switching-voltage ferroelectric ($La_xBi_{1-x}FeO_3$)

© 2019 Springer Nature Limited. All rights reserved.

Table 2 | Material options for MESO logic

	Type 1	Type 2	Type 3
SOC materials for spin-to-charge conversion	High-SOC and topological oxides	Topological materials and superlattices	Two-dimensional transition-metal dichalcogenides
	Bi_2O_3[44], $SrIrO_3$[45], $SrTiO_3/LaAlO_3$[25,36,38]	$Bi_{1.5}Sb_{0.5}Te_{1.7}Se_{1.3}$[24], Bi_2Se_3[34,35], α-Sn[46], BiSb[47]	MoS_2[48], MX_2[49]
Magneto-electrics	Multiferroics	Magnetostrictive	Exchange bias
	$BiFeO_3$[27,30], $LaBiFeO_3$[42], $TbMnO_3$[50], $LuFeO_3/LuFe_2O_4$[51]	Fe_3Ga[52], $Tb_xDy_{1-x}Fe_2$[53], FeRh[28]	Cr_2O_3[29,54], Fe_2TeO_6[55]
Interconnect	Noble metals	Metal–semiconductor	Interlayer dielectric
	Cu, Ag, Co, Al, Ru	poly-Si, NiSi, CoSi, NiGe, TiSi	SiO_2, SiN, SiCOH, polymers
Nanomagnet	Nominal ferromagnets	Heusler alloys	
	Co, Fe, Ni, CoFe, NiFe	X_2YZ and XYZ alloys (for example, Co_2FeAl, Mn_3Ga)	

Three classes of materials (high-SOC oxides, topological materials and superlattices, and two-dimensional transition-metal dichalcogenides) are suitable for SOC-based spin-to-charge conversion. Magnetoelectrics belong to three classes: (1) multiferroics with magnetic (antiferromagnetic/ferromagnetic) and electric (ferroelectric) order parameters, (2) magnetostrictive, that is, one ferromagnetic-order-parameter material combined with a strain/piezoelectric material, and (3) exchange-bias materials, that is, one magnetic-order parameter with no ferroelectric/antiferroelectric order. Magnetostrictive materials are not directly suitable (because only 90° switching is feasible), but can be used to enhance magnetoelectric switching. Interconnect options comprise noble metals, metal–semiconductors (which exhibit excellent gap fill for interconnect processing and have short electron mean free paths) and interlayer dielectrics chosen for their low dielectric constant. Nanomagnets should be conductive to allow spin injection with the applied bias. Co-, Fe- and Ni-based ferromagnets or Heusler alloys are potential candidates, with low M_s and high spin polarization.

was obtained using symmetric conductive-oxide electrodes ($SrRuO_3$), from which the switching energy was estimated to be about1 aJ per bit for a contact area of 10 nm × 10 nm.

Material requirements for 1–10-aJ-class MESO logic

We describe the material scaling requirements for 1–10-aJ-class MESO logic scalable to critical dimensions of <10 nm or device density beyond 10^{10} cm^{-2}. In Table 1 we summarize the material scaling requirements for four classes of materials: a) SOC materials for spin-to-charge conversion, (2) magnetoelectrics for charge-to-spin conversion, (3) interconnects scalable to nanoscale widths and (4) nanomagnets. We considered the experimental values shown for the inverse Rashba–Edelstein parameters. A large-signal magnetoelectric coefficient of $10c^{-1}$ (c, speed of light) from magnetoelectric switching and low coercive voltages were obtained via rhombohedral distortion tuning and chemical substitution of multiferroics[42] with thickness scalability to 5–20 nm. The output resistance of the ISOC spin current source is a critical parameter that affects the driving ability of the MESO logic device (high source resistance is preferred for a current source; see Supplementary Information section H and Supplementary Fig. 9.) The requirement of low interconnect resistivity is considerably relaxed owing to the low-voltage, low-current operation of charge-mediated magnetoelectric logic. This is reflected in the resistivity target of 4–200 μΩ cm, which is comparable to the resistivity of scaled metal wires. Electromigration of the metal interconnect imposes a challenging limit on the switching speed by limiting the peak current in wires. MESO logic relaxes the electromigration requirements to 25 MA cm^{-2}, appreciably below the Belch limit for electromigration of interconnect metal candidates[43].

A focused effort using quantum materials can enable logic technology operating at 100 mV and 1 aJ per bit. The details of four fundamental material classes are presented in Table 2. SOC materials can be comprised of (1) high-SOC oxides (for example, W(O) and Bi_2O_3[44]) and oxides with strong topological effects ($SrIrO_3$[45] and $SrTiO_3/LaAlO_3$[25,36,38]), (2) topological materials ($Bi_{1.5}Sb_{0.5}Te_{1.7}Se_{1.3}$[24], Sn-$Bi_2Te_2Se$, Bi_2Se_3[34,35], α-Sn[46], BiSb[47]) and their superlattices and (3) transition-metal dichalcogenides with large spin–orbit effects (MoS_2[48], MX_2[49]). The magnetoelectric materials can be comprised of (1) multiferroics with coupling of the antiferromagnetic and ferroelectric orders (type -1 multiferroics $BiFeO_3$[27,30] and $LaBiFeO_3$[42]; type-2 multiferroics, such as $TbMnO_3$[50]; and improper multiferroics, such as $LuFeO_3/LuFe_2O_4$[51]), (2) magnetostrictive materials (Fe_3Ga[52], $Tb_xDy_{1-x}Fe_2$[53], FeRh[18]) and (3) electrically tuned exchange-mediated magnetoelectrics (Cr_2O_3[29,54] or Fe_2TeO_6[55]). The interconnect options scalable to dimensions smaller than the 10 nm critical width can be based on transition metals (Cu, Ag, Co, Al, Ru) or their semiconductor alloys (poly-Si, NiSi, CoSi, NiGe, TiSi), combined with low-interconnect-capacitance materials (SiO_2, SiN, SiCOH, polymers). The nanomagnetic materials can be ferromagnets/ferrimagnets (Co, Fe, Ni, CoFe, NiFe, and X_2YZ and XYZ alloys, such as Co_2FeAl, Mn_3Ga), in which a wide range of saturation magnetization and magnetic anisotropy are feasible to meet the dimensionality and retention requirements. In each of these four classes of materials, considerable development is required to improve the material interfaces for integrated devices, the operating temperature range, the processing temperature compatibility and, most importantly, the performance metrics.

Conclusion

In conclusion, we propose a scalable beyond-CMOS spintronic logic device with non-volatility and an energy-efficient charge-based interconnect. The proposed device allows (a) continued scaling in energy per operation towards attojoule-level switching energy (about 30 times below that of advanced CMOS devices) at 100 mV (more than 5 times below the operating voltage of advanced CMOS devices), (2) substantial improvement in logic density (about 5 times compared to advanced CMOS devices), enabled by majority-gate circuits implemented with a collective switching device, (3) improved scalability for interconnects due to the small impact of the resistivity, which is up to 1 mΩ cm, and (4) a path to seamless monolithic integration with CMOS technology (see Supplementary Fig. 19). The development of a beyond-CMOS device with an advantageous scaling method using quantum materials, highly compact majority logic[14] and non-volatile logic[15] can open up a potentially new technology paradigm for improving energy efficiency in beyond-CMOS computing devices. Combined with non-volatility and ultra-low energy, MESO logic may enable entirely new computer architectures that may avoid the trade-offs of the Turing and von Neumann architectures and of Amdahl's law. A combination of quantum materials, novel integration and new logic architectures may thus enable computing beyond advanced CMOS technology.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0770-2.

Received: 24 April 2016; Accepted: 7 October 2018;
Published online 3 December 2018.

1. Kuhn, K. J. Considerations for ultimate CMOS scaling. *IEEE Trans. Electron Dev.* **59**, 1813–1828 (2012).
2. Ferain, I., Colinge, C. A. & Colinge, J.-P. Multigate transistors as the future of classical metal-oxide-semiconductor field-effect transistors. *Nature* **479**, 310–316 (2011).

© 2019 Springer Nature Limited. All rights reserved.

3. Auth, C. et al. A 10 nm high performance and low-power CMOS technology featuring 3rd generation FinFET transistors, self-aligned quad patterning, contact over active gate and cobalt local interconnects. In *Electron Devices Meeting 2017*, 29.1.1–29.1.4 (IEEE, 2017).
4. Moore, G. E. Cramming more components onto integrated circuits. *Proc. IEEE* **86**, 82–85 (1998).
5. Dennard, R. H., Gaensslen, F. H., Yu, H. N., Rideout, V. L., Bassous, E. & Leblanc, A. R. Design of ion-implanted MOSFET's with very small physical dimensions. *IEEE J. Solid St. Circ.* **9**, 256–268 (1974).
6. Ghani, T. et al. A 90 nm high volume manufacturing logic technology featuring novel 45nm gate length strained silicon CMOS transistors. In *Electron Devices Meeting 2003*, 11.6.1–11.6.3 (IEEE, 2003).
7. Krishnamohan, T. et al. Comparison of (001), (110) and (111) uniaxial- and biaxial- strained-Ge and strained-Si PMOS DGFETs for all channel orientations: mobility enhancement, drive current, delay and off-state leakage. In *Electron Devices Meeting 2008*, 1–4 (IEEE, 2008).
8. Huang, X. et al. Sub 50-nm FinFET: PMOS. In *Electron Devices Meeting 1998*, 67–70 (IEEE, 1999).
9. Schumacher, M., Baumann, P. K. & Seidel, T. AVD and ALD as two complementary technology solutions for next generation dielectric and conductive thin-film processing. *Chem. Vap. Depos.* **12**, 99–108 (2006).
10. Horowitz, M. Computing's energy problem (and what we can do about it). In *Solid-State Circuits Conference Digest of Technical Papers 2014* 10–14 (IEEE, 2014).
11. Theis, T. N. & Solomon, P. M. It's time to reinvent the transistor! *Science* **327**, 1600–1601 (2010).
12. Nikonov, D. E. & Young, I. A., Benchmarking of beyond-CMOS exploratory devices for logic integrated circuits. *IEEE J. Explor. Solid-State Computat. Devices Circuits* **1**, 3–11 (2015).
13. Manipatruni, S., Nikonov, D. E. & Young, I. A. Beyond CMOS computing with spin and polarization. *Nat. Phys.* **14**, 338 (2018).
14. Zografos, O. et al. Design and benchmarking of hybrid CMOS-spin wave device circuits compared to 10 nm CMOS. In *2015 IEEE 15th International Conference on Nanotechnology (IEEE-NANO)*, 686–689 (IEEE, 2015).
15. Ma, K. et al. Nonvolatile processor architecture exploration for energy-harvesting applications. *IEEE Micro* **35**, 32–40 (2015).
16. Mead, C. Neuromorphic electronic systems. *Proc. IEEE* **78**, 1629–1636 (1990).
17. Patil, A. D., Manipatruni, S., Nikonov, D., Young, I. A. & Shanbhag, N. R. 2017. Shannon-inspired statistical computing to enable spintronics. Preprint at https://arxiv.org/abs/1702.06119 (2017).
18. Dyakonov, M. I. & Perel, V. I. Current-induced spin orientation of electrons in semiconductors. *Phys. Lett. A* **35**, 459 (1971).
19. Edelstein, V. M. Spin polarization of conduction electrons induced by electric current in two-dimensional asymmetric electron systems. *Solid State Commun.* **73**, 233–235 (1990).
20. Soumyanarayanan, A., Reyren, N., Fert, A. & Panagopoulos, C. Emergent phenomena induced by spin–orbit coupling at surfaces and interfaces. *Nature* **539**, 509 (2016).
21. Hsieh, D. et al. A tunable topological insulator in the spin helical Dirac transport regime. *Nature* **460**, 1101 (2009).
22. Rojas Sánchez, J. C. et al. Spin-to-charge conversion using Rashba coupling at the interface between non-magnetic materials. *Nat. Commun.* **4**, 2944 (2013).
23. Shen, K., Vignale, G. & Raimondi, R. Microscopic theory of the inverse Edelstein effect. *Phys. Rev. Lett.* **112**, 096601 (2014).
24. Shiomi, Y. et al. Spin–electricity conversion induced by spin injection into topological insulators. *Phys. Rev. Lett.* **113**, 196601 (2014).
25. Varignon, J., Vila, L., Barthelemy, A. & Bibes, M. A new spin for oxide interfaces. *Nat. Phys.* **14**, 322 (2018).
26. Spaldin, N. A. & Fiebig, M. The renaissance of magnetoelectric multiferroics. *Science* **309**, 391–392 (2005).
27. Heron, J. T. et al. Deterministic switching of ferromagnetism at room temperature using an electric field. *Nature* **516**, 370–373 (2014).
28. Cherifi, R. O. et al. Electric-field control of magnetic order above room temperature. *Nat. Mater.* **13**, 345–351 (2014).
29. He, X. et al. Robust isothermal electric control of exchange bias at room temperature. *Nat. Mater.* **9**, 579–585 (2010).
30. Manipatruni, S. et al. Voltage control of uni-directional anisotropy in ferromagnet–multiferroic system. Preprint at https://arxiv.org/abs/1801.08280 (2018).
31. Brataas, A., Bauer, G. E. & Kelly, P. J. Non-collinear magnetoelectronics. *Phys. Rep.* **427**, 157–255 (2006).
32. Manipatruni, S., Nikonov, D. E. & Young, I. A. Modeling and design of spintronic integrated circuits. *IEEE Trans. Circuits Syst.* **59**, 2801–2814 (2012).
33. Omori, Y. et al. Inverse spin Hall effect in a closed loop circuit. *Appl. Phys. Lett.* **104**, 242415 (2014).
34. Fan, Y. et al. Magnetization switching through giant spin–orbit torque in a magnetically doped topological insulator heterostructure. *Nat. Mater.* **13**, 699 (2014).
35. Mahendra, D. C. et al. Room-temperature perpendicular magnetization switching through giant spin-orbit torque from sputtered $Bi_xSe(1-x)$ topological insulator material. Preprint at https://arxiv.org/abs/1703.03822 (2017).
36. Lesne, E. et al. Highly efficient and tunable spin-to-charge conversion through Rashba coupling at oxide interfaces. *Nat. Mater.* **15**, 1261–1266 (2016).
37. Ast, C. R. et al. Giant spin splitting through surface alloying. *Phys. Rev. Lett.* **98**, 186807 (2007).
38. Veit, M. J., Arras, R., Ramshaw, B. J., Pentcheva, R. & Suzuki, Y. Nonzero Berry phase in quantum oscillations from giant Rashba-type spin splitting in $LaTiO_3$/$SrTiO_3$ heterostructures. *Nat. Commun.* **9**, 1458 (2018).
39. Meindl, J. D., Chen, Q. & Davis, J. A. Limits on silicon nanoelectronics for terascale integration. *Science* **293**, 2044–2049 (2001).
40. Manipatruni, S., Lipson, M. & Young, I. A. Device scaling considerations for nanophotonic CMOS global interconnects. *IEEE J. Sel. Topics Quantum Electron.* **19**, 8200109 (2013).
41. Mayadas, A. F., Shatzkes, M. & Janak, J. F. Electrical resistivity model for polycrystalline films: the case of specular reflection at external surfaces. *Appl. Phys. Lett.* **14**, 345–347 (1969).
42. Chu, Y. H. et al. Low voltage performance of epitaxial $BiFeO_3$ films on Si substrates through lanthanum substitution. *Appl. Phys. Lett.* **92**, 102909 (2008).
43. Gardner, D. S., Meindl, J. D. & Saraswat, K. C. Interconnection and electro migration scaling theory. *IEEE Trans. Electron Dev.* **34**, 633–643 (1987).
44. Karube, S., Kondou, K. & Otani, Y. 2016. Experimental observation of spin to charge current conversion at non-magnetic metal/Bi_2O_3 interfaces. Preprint at https://arxiv.org/abs/1601.04292 (2016).
45. Pesin, D. & Balents, L. Mott physics and band topology in materials with strong spin–orbit interaction. *Nat. Phys.* **6**, 376 (2010).
46. Rojas-Sánchez, J.-C. et al. Spin to charge conversion at room temperature by spin pumping into a new type of topological insulator: α-Sn films. *Phys. Rev. Lett.* **116**, 096602 (2016).
47. Khang, N. H. D., Ueda, Y. & Hai, P. N. A conductive topological insulator with large spin Hall effect for ultralow power spin–orbit torque switching. *Nat. Mater.* **17**, 808–813 (2018).
48. Cheng, C. et al. Direct observation of spin-to-charge conversion in MoS_2 monolayer with spin pumping. Preprint at https://arxiv.org/abs/1510.03451 (2015).
49. Wang, G. et al. Spin–orbit engineering in transition metal dichalcogenide alloy monolayers. *Nat. Commun.* **6**, 10110 (2015).
50. Kimura, T., Goto, T., Shintani, H., Ishizaka, K., Arima, T. & Tokura, Y. Magnetic control of ferroelectric polarization. *Nature* **426**, 55–58 (2003).
51. Mundy, J. A. et al. Atomically engineered ferroic layers yield a room-temperature magnetoelectric multiferroic. *Nature* **537**, 523 (2016).
52. Srisukhumbowornchai, N. & Guruswamy, S. Large magnetostriction in directionally solidified FeGa and FeGaAl alloys. *J. Appl. Phys.* **90**, 5680–5688 (2001).
53. Ryu, J., et al. Magnetoelectric properties in piezoelectric and magnetostrictive laminate composites. *Jpn J. Appl. Phys.* **40**, 4948 –4951 (2001).
54. Street, M. et al. Increasing the Néel temperature of magnetoelectric chromia for voltage-controlled spintronics. *Appl. Phys. Lett.* **104**, 222402 (2014).
55. Wang, J. et al. Magnetoelectric Fe_2TeO_6 thin films. *J. Phys. Condens. Matter* **26**, 055012 (2014).

Acknowledgements We are grateful to A. Fert and J.-P. Wang for discussions. We acknowledge F. Rana, D. Schlom and F. Casanova for insights shared with us. We also acknowledge the support of K. Oguz and B. Buford of Intel Corporation for discussions on device integration and metrology. R.R. acknowledges the long-term support of the Quantum Materials programme funded by the US Department of Energy, Office of Basic Energy Sciences, which laid the foundation for the key elements of the work reported in this paper. B.P., Y.-L.H. and R.R. acknowledge support from Semiconductor Research Corporation within the JUMP program.

Reviewer information *Nature* thanks V. Bertacco, Y. Otani and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions S.M. identified the use of the inverse spin–orbit effect for electrically transduced spin-logic devices. S.M., D.E.N. and I.A.Y. developed the logic circuits. S.M. developed the scaling laws, physical macro models and interconnect estimates. H.L. developed circuit design techniques and performed the logic-circuit simulations with the physical macro models. S.M. and R.R. developed the material scaling options and coordinated the material growth. S.M. conceptualized the test devices and designed the experiments and measurements for the magnetoelectric and spin–orbit devices. D.E.N. benchmarked the performance of the circuits. C.-C.L., B.P. and T.G. performed the layout of the test devices, processed the devices and identified processing methods for sub-micron-sized magnetoelectric and spin–orbit devices. S.M. and E.B. measured the magnetoelectric devices. S.M. and T.G. measured the spin–orbit devices. B.P., Y.-L.H. and E.B. deposited the samples and performed material characterization under the supervision of R.R. S.M. wrote the manuscript and D.E.N., I.A.Y. and R.R. edited the manuscript. All authors reviewed the manuscript and interpreted the data.

Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0770-2.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to S.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

Uniform benchmarking to beyond-CMOS logic options. We adopted the uniform, beyond-CMOS benchmarking method[12] to compare beyond-CMOS options. This method describes the impact of material improvements, device parameters, circuit topology and interconnects on the performance of computing devices. The model is adopted in beyond-CMOS research and includes the following spin-logic devices: (1) spin-torque devices[32,56–58] (spin-transfer-torque domain-wall device, all-spin-logic device, charge spin logic, spin-torque oscillator logic), (2) dipole-field devices (nanomagnetic logic)[59], and (3) magnetoelectric devices (MESO, spin majority gate, spin-wave device[60]). We evaluated several digital logic circuits[12] (a fanout-4 inverter, a two-input NAND adder, a 32-bit ripple-carry adder and a 32-bit ALU) to compare the MESO logic with leading beyond-CMOS logic options. We also considered tunnelling field-effect transistors[61,62], ferroelectric and piezoelectric integration in transistors[63,64] and Mott transistors[65]. See Supplementary Information section F, for a detailed explanation of the benchmarking.

Vector spin circuit modelling of MESO logic. We verify the functionality of MESO spin logic using an equivalent spin circuit model that describes the magnetization dynamics of the nanomagnet, vector spin injection, spin-to-charge transduction and magnetoelectric switching. The equivalent circuit model is based on vector spin circuit theory (magnetoelectric circuit analysis); see Supplementary Information section B.

The intrinsic resistance of the ISOC current source is derived from the conductivity of the interconnect and the ISOC conversion layers. The nanomagnet is connected to a control transistor operating as a power supply and shared among several MESO devices. We have also included the resistance and capacitance parasitics of the ground contact. The conductance across the magnet to the spin injection layer is modelled as a 4×4 matrix that relates the four-component charge and spin voltages to the injected four-component charge and spin currents[31,32,66]. The current injected at the nanomagnet–ferromagnet interface is given by

$$\begin{bmatrix} I_c \\ I_{sx} \\ I_{sy} \\ I_{sz} \end{bmatrix} = R^{-1}(\hat{\boldsymbol{m}}) \begin{bmatrix} G_{11} & \alpha G_{11} & 0 & 0 \\ \alpha G_{11} & G_{11} & 0 & 0 \\ 0 & 0 & G_{\mathrm{SL}} & G_{\mathrm{FL}} \\ 0 & 0 & -G_{\mathrm{FL}} & G_{\mathrm{SL}} \end{bmatrix} R(\hat{\boldsymbol{m}}) \begin{bmatrix} V_{\mathrm{N}} - V_{\mathrm{F}} \\ V_{sx} \\ V_{sy} \\ V_{sz} \end{bmatrix} \quad (3)$$

where R is the rotation matrix that accounts for the magnetization direction of the nanomagnet, and I_{si} and V_{si}, $i = \{x, y, z\}$, are the components of the spin current and voltage, respectively. G_{SL}, G_{FL}, G_{11}, V_{N} and V_{F} are the conductance elements for the Slonczewski torque, field-like torque, conductance, voltage at the normal metal and voltage at the ferromagnet, correspondingly. See Supplementary Information sections B and C for a detailed explanation of the model.

Stochastic behaviour of magnetoelectric switching versus spin-torque switching. We modelled the magnetization dynamics of the nanomagnet using the Landau–Lifshitz–Gilbert equation[50] and the Fokker–Planck equation[67,68]. The modified Landau–Lifshitz–Gilbert equation, a phenomenological equation that describes the dynamics of a nanomagnet with magnetic moment unit vector $\hat{\boldsymbol{m}}$, was used for Monte Carlo simulations (see Supplementary Table 1 for parameters). We used the Fokker–Planck equation for uniaxial anisotropy parameterized with the angle of the magnetization, which was validated versus the Monte Carlo simulations of the nanomagnets. See Supplementary Information sections C and O for a detailed explanation of the stochastic modelling.

Complete logic family and state elements. The proposed device family is readily extended to a general-purpose computing state machine. A state machine and complete Boolean logic family are the prerequisites for a Turing machine[69]. Majority logic operation can be readily demonstrated because the input of a capacitive node is added to the charge currents converging at the node via the Kirchhoff law. Spin-logic devices with multiple switching inputs (domain-wall, spin-wave or spin-current) have been shown to facilitate the development of majority logic[70] and spin state machines[71]. Combined with a random access memory (RAM), a state machine enables general-purpose computing.

CMOS compatibility, memory and control logic. The proposed magnetoelectric-logic scalable spintronic logic device has several desirable features that are compatible with CMOS nanoelectronics. First, the MESO device can be integrated in the backend of the CMOS process (that is, between the interconnect layers), allowing CMOS devices to be used for clocking, control and power supply (Supplementary Fig. 19). Second, MESO devices can serve as elements of an embedded memory with 'logic-compatible speed' (known as large-signal memory, commonly implemented with a static RAM), making them usable as on-chip non-volatile memories. Third, the MESO device allows stacking of several layers of magnetic logic in a three-dimensional architecture. Fourth, because the state variable of the interconnect between MESO gates is the charge, MESO logic can be readily interfaced with CMOS circuitry to implement clocking control and power delivery. Owing to its low supply voltage, the MESO device is efficient even with interconnects with high metal resistivity of $>100\ \mu\Omega\ \mathrm{cm}$[72,73].

Code availability. The MATLAB codes used to benchmark the circuit performance are available under 'Benchmarking of devices' from the Nanoelectronics Research Initiative, at https://nanohub.org/tools/nribench/browser/trunk/src.

Data availability

The data that support the findings of this study are available from the corresponding author on reasonable request.

56. Behin-Aein, B., Datta, D., Salahuddin, S. & Datta, S. Proposal for an all-spin logic device with built-in memory. *Nat. Nanotechnol.* **5**, 266–270 (2010).
57. Nikonov, D. E., Bourianoff, G. I. & Ghani, T. Proposal of a spin torque majority gate logic. *IEEE Electron Device Lett.* **32**, 1128–1130 (2011).
58. Manipatruni, S., Nikonov, D. E. & Young, I. A. Material targets for scaling all-spin logic. *Phys. Rev. Appl.* **5**, 014002 (2016).
59. Imre, A. et al. Majority logic gate for magnetic quantum-dot cellular automata. *Science* **311**, 205–208 (2006).
60. Chumak, A. V., Serga, A. A. & Hillebrands, B. Magnon transistor for all-magnon data processing. *Nat. Commun.* **5**, 4700 (2014).
61. Ionescu, A. M. & Riel, H. Tunnel field-effect transistors as energy-efficient electronic switches. *Nature* **479**, 329–337 (2011).
62. Lu, H. & Seabaugh, A. Tunnel field-effect transistors: state-of-the-art. *IEEE J. Electron Devices Soc.* **2**, 44–49 (2014).
63. Salahuddin, S. & Datta, S. Use of negative capacitance to provide voltage amplification for low power nanoscale devices. *Nano Lett.* **8**, 405–410 (2008).
64. Newns, D., Elmegreen, B., Liu, X. H. & Martyna, G. A low-voltage high-speed electronic switch based on piezoelectric transduction. *J. Appl. Phys.* **111**, 084509 (2012).
65. Son, J., Rajan, S., Stemmer, S. & Allen, S. J. A heterojunction modulation-doped Mott transistor. *J. Appl. Phys.* **110**, 084503 (2011).
66. Srinivasan, S., Diep, V., Behin-Aein, B., Sarkar, A. & Datta, S. Modeling multi-magnet networks interacting via spin currents. In *Handbook of Spintronics* 1281–1335 (2016).
67. Apalkov, D. M. & Visscher, P. B. Spin-torque switching: Fokker–Planck rate calculation. *Phys. Rev. B* **72**, 180405 (2005).
68. Butler, W. H. et al. Switching distributions for perpendicular spin-torque devices within the macrospin approximation. *IEEE Trans. Magnet.* **48**, 4684–4700 (2012).
69. Shannon, C. E. A universal Turing machine with two internal states. *Automata Stud.* **34**, 157–165 (1957).
70. Amarù, L. et al. Majority logic synthesis. In *Proc. International Conference on Computer-Aided Design* 79 (ACM, 2018).
71. Manipatruni, S., Nikonov, D. E. & Young, I. A. All-spin nanomagnetic state elements. *Appl. Phys. Lett.* **103**, 063503 (2013).
72. Dutta, S. et al. Highly scaled ruthenium interconnects. *IEEE Electron Dev. Lett.* **38**, 949–951 (2017).
73. Dutta, S. et al. Sub-100 nm^2 cobalt interconnects. *IEEE Electron Dev. Lett.* **39**, 731–734 (2018).

© 2019 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0768-9

Loss of ADAR1 in tumours overcomes resistance to immune checkpoint blockade

Jeffrey J. Ishizuka[1,2,3,9], Robert T. Manguso[1,3,9], Collins K. Cheruiyot[1,3], Kevin Bi[1,3], Arpit Panda[1,3,4], Arvin Iracheta-Vellve[1,3], Brian C. Miller[1,2,3], Peter P. Du[1,3], Kathleen B. Yates[1,3], Juan Dubrot[1,3], Ilana Buchumenski[5], Dawn E. Comstock[1,3,4], Flavian D. Brown[1,3,4], Austin Ayer[1,3], Ian C. Kohnle[1,3], Hans W. Pope[1,3], Margaret D. Zimmer[1,3], Debattama R. Sen[1,3,4], Sarah K. Lane-Reticker[1,3], Emily J. Robitschek[1,3], Gabriel K. Griffin[1,3,6], Natalie B. Collins[1,3,7], Adrienne H. Long[1,3], John G. Doench[3], David Kozono[8], Erez Y. Levanon[5] & W. Nicholas Haining[1,3,7]*

Most patients with cancer either do not respond to immune checkpoint blockade or develop resistance to it, often because of acquired mutations that impair antigen presentation. Here we show that loss of function of the RNA-editing enzyme ADAR1 in tumour cells profoundly sensitizes tumours to immunotherapy and overcomes resistance to checkpoint blockade. In the absence of ADAR1, A-to-I editing of interferon-inducible RNA species is reduced, leading to double-stranded RNA ligand sensing by PKR and MDA5; this results in growth inhibition and tumour inflammation, respectively. Loss of ADAR1 overcomes resistance to PD-1 checkpoint blockade caused by inactivation of antigen presentation by tumour cells. Thus, effective anti-tumour immunity is constrained by inhibitory checkpoints such as ADAR1 that limit the sensing of innate ligands. The induction of sufficient inflammation in tumours that are sensitized to interferon can bypass the therapeutic requirement for CD8$^+$ T cell recognition of cancer cells and may provide a general strategy to overcome immunotherapy resistance.

Despite the remarkable clinical successes of immune checkpoint blockade, most patients do not respond to immunotherapy, or develop therapeutic resistance because of mutations in the interferon-γ (IFNγ)-sensing pathway or in the antigen-presentation pathway[1–4]. There are currently no therapeutic options to overcome acquired resistance to checkpoint blockade.

We recently conducted a pooled in vivo CRISPR screen to identify genes expressed by the B16 transplantable melanoma model that, when deleted, confer sensitivity to immunotherapy[5]. This screen identified a number of genes with the potential to modify the response to endogenous RNA species, including *Adar1*, which encodes an adenosine deaminase that binds to and limits the sensing of endogenous double-stranded RNA (dsRNA)[6–9]. Here we show that ADAR1 functions as a checkpoint that limits anti-tumour immunity by preventing the sensing of endogenous dsRNA. Loss of function of ADAR1 improves responses to PD-1 blockade and overcomes common mechanisms of resistance to immunotherapy.

ADAR1 loss sensitizes tumours to immunotherapy

In our prior in vivo CRISPR screen, *Adar1*-targeting single-guide RNAs (sgRNAs) were markedly depleted from tumours in immunocompetent mice (Fig. 1a). To test whether deletion of *Adar1* sensitized tumours to anti-tumour immunity, we generated mouse B16 tumour cells that lacked ADAR1 (Extended Data Fig. 1a) and compared their growth with control B16 tumours in vitro and in vivo. B16 cells lacking ADAR1 p150 or ADAR1 p110/p150 (each isoform targeted by three sgRNAs; hereafter termed *Adar1*-null tumours) grew equivalently to control tumour cells in vitro (Fig. 1b), suggesting that neither isoform of ADAR1 is essential for cell growth. *Adar1*-null tumours implanted in immunodeficient NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ (NSG) mice showed only a minimal decrease in tumour size (Fig. 1c). By contrast, *Adar1*-null B16 tumours in wild-type, immunocompetent animals were profoundly sensitized to anti-PD-1 antibody treatment ($P < 0.0001$, log-rank test; Fig. 1c, Extended Data Fig. 1a). We found similar results in *Adar1*-null tumour cell lines in three additional transplantable tumour models and also when tumours were size-matched at the time of treatment (Extended Data Fig. 1c–e, Supplementary Text I). Thus, *Adar1*-null tumours are sensitized to anti-tumour immunity and immunotherapy in multiple transplantable tumour models.

Loss of ADAR1 increases tumour inflammation

We compared the immune microenvironment of *Adar1*-null and control B16 tumours from untreated wild-type mice and found a significant increase in CD8$^+$ T cells in *Adar1*-null tumours ($P < 0.005$, Student's *t*-test; Fig. 2a) that were infiltrated throughout the tumour, significantly increased CD45$^+$ immune cell infiltration ($P < 0.01$, Student's *t*-test; Fig. 2b, Extended Data Figs. 2, 3) and significantly increased proportions of CD3$^+$ T cells, CD4$^+$ T cells, CD8$^+$ T cells, γδ T cells and natural killer (NK) cells ($P < 0.0001$, $P < 0.05$, $P < 0.0001$, $P < 0.001$ and $P < 0.05$, respectively, Student's *t*-test; Fig. 2b). *Adar1*-null tumours had significantly decreased proportions of myeloid-derived suppressor cells (MDSCs) and tumour-associated neutrophils ($P < 0.01$ and $P < 0.05$, respectively, Student's *t*-test; Fig. 2b).

Single-cell RNA sequencing of CD45$^+$ cells in the tumour microenvironment (TME) (Fig. 2c, Extended Data Fig. 4a, b) confirmed increased CD8$^+$ T cell infiltration and showed a striking repolarization of the myeloid compartment of *Adar1*-null tumours (Fig. 2d). M2 macrophages and MDSCs were decreased in *Adar1*-null tumours relative to control tumours (Fig. 2d), and there was a marked decrease in expression of genes in myeloid cells associated with a suppressive phenotype (Extended Data Fig. 4c, Wilcoxon

[1]Department of Pediatric Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [2]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [3]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Division of Medical Sciences, Harvard Medical School, Boston, MA, USA. [5]Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat Gan, Israel. [6]Department of Pathology, Brigham and Women's Hospital, Boston, MA, USA. [7]Division of Pediatric Hematology and Oncology, Children's Hospital, Boston, Massachusetts, USA. [8]Department of Radiation Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [9]These authors contributed equally: Jeffrey J. Ishizuka, Robert T. Manguso.
*e-mail: nicholas_haining@dfci.harvard.edu

© 2019 Springer Nature Limited. All rights reserved.

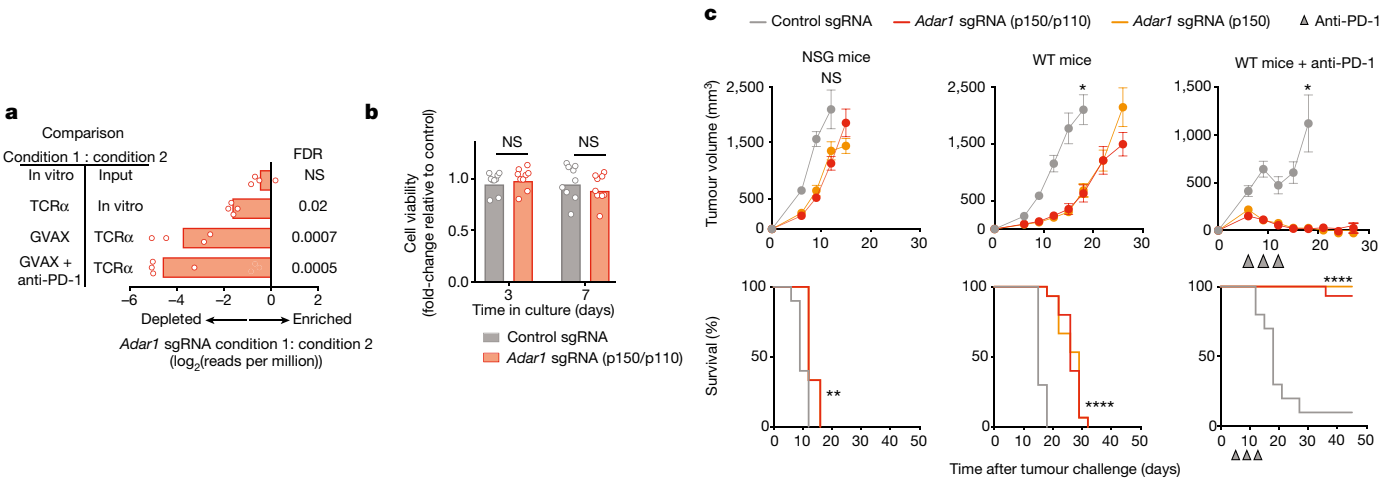


Fig. 1 | Loss of ADAR1 in tumour cells enhances anti-tumour immunity and responses to PD-1 checkpoint blockade. a, Relative depletion of *Adar1* sgRNAs from a pool of sgRNAs targeting 2,368 genes expressed by Cas9$^+$ B16 tumour cells. $n = 4$ independent guides targeting each gene; false discovery rate (FDR) was calculated using the STARS algorithm v1.3 to generate permutation testing and a null distribution. **b**, Viability of *Adar1*-null and control B16 tumour cells following in vitro culture. $n = 9$ for each condition; results are representative of two independent experiments. **c**, Tumour volume and survival analysis of control (grey), *Adar1* p150-null (orange) or *Adar1* p110/p150-null (red) B16 tumours in NSG, wild-type (WT) and wild-type anti-PD-1-treated C57BL/6 mice. Wild-type B16 mice that did not receive anti-PD-1 treatment were treated with an equivalent concentration of rat IgG2a isotype control antibody. Data in **c** represent two independent experiments with $n = 5$ animals per guide with two separate guides for the control group and three separate guides for each *Adar1*-null group. Data for individual guides targeting each isoform of ADAR1 are pooled. **a**, **b**, Bars show mean; **c**, mean ± s.e.m. **b**, **c** (tumour volume curves), two-sided Student's *t*-test; **c** (survival curves), log-rank test; $^*P < 0.05$; $^{**}P < 0.01$; $^{****}P < 0.0001$; NS, not significant.

rank-sum test). We found a shift in the balance of chemokines expressed by immune cells in *Adar1*-null tumours, with decreased expression of chemokines associated with the recruitment of MDSCs and increased expression of chemokines associated with the recruitment of T cells and NK cells (Extended Data Fig. 4c). Genes associated with the activation and effector function of CD8$^+$ T cells were increased in *Adar1*-null tumours relative to controls (Extended Data Fig. 4c).

We found an increase in the expression of gene signatures for IFNα and IFNγ responses in almost all types of immune cell from *Adar1*-null tumours relative to control tumours (Fig. 2e, Extended Data Fig. 4d) and both IFNβ and IFNγ protein levels in tumour lysates were significantly higher in *Adar1*-null tumours than in control tumours (Fig. 2f, $P < 0.01$ for both, Student's *t*-test). Deletion of *Adar1* therefore causes a global reshaping of the tumour immune compartment and increased abundance of IFNs.

Loss of ADAR1 sensitizes tumours to IFNs

Adar1-null tumours were significantly more sensitive than control tumours to T cell killing ($P < 0.0001$, $P < 0.05$, $P < 0.01$ for

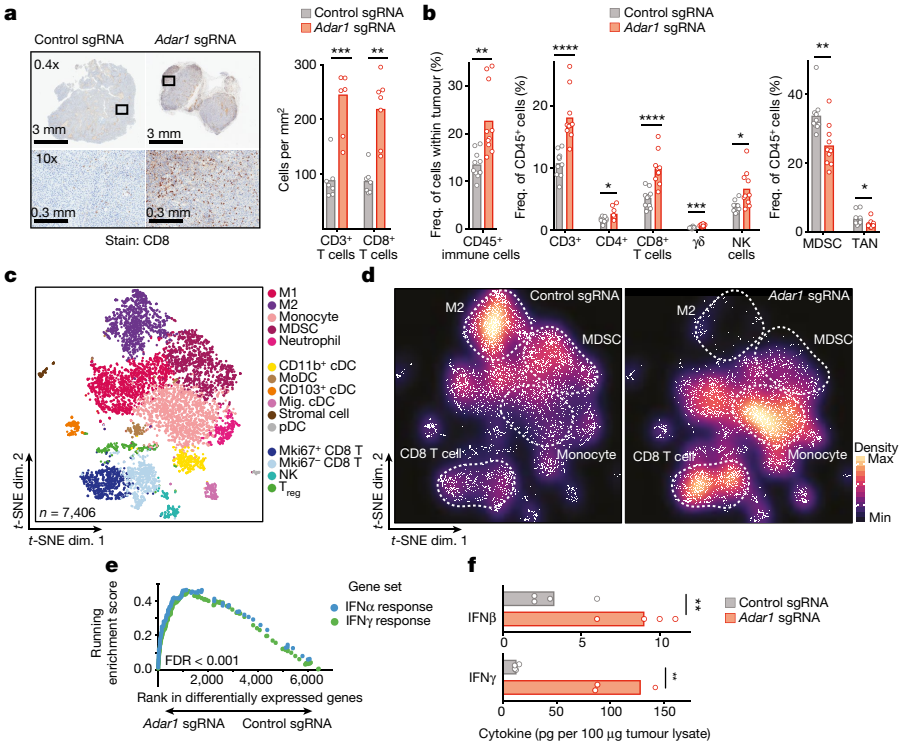


Fig. 2 | Loss of ADAR1 inflames the TME. a, Immunohistochemistry (left) with quantification (right) of CD3$^+$ and CD8$^+$ T cells in untreated control (grey) or *Adar1*-null (red) B16 tumours ($n = 8$ mice per group). **b**, Flow cytometry of immune populations from untreated control and *Adar1*-null B16 tumours (representative results from two independent experiments; $n = 8$–10 mice per group per experiment). TAN, tumour-associated neutrophils. **c**, **d**, *t*-Distributed stochastic neighbour embedding (*t*-SNE) plot (**c**) and density plots (**d**) of 7,406 RNA-sequenced (RNA-seq) single CD45$^+$ cells from *Adar1*-null and control B16 tumours ($n = 2$ biological replicates for each population). T_{reg}, regulatory T cells; M1, M1 macrophages; M2, M2 macrophages; cDC, conventional dendritic cells; MoDC, monocyte-derived dendritic cells; pDC, plasmacytoid dendritic cells. **e**, Gene set enrichment analysis (GSEA) of IFN response signatures in immune cells from *Adar1*-null compared with control tumours. FDR calculated using GSEA. **f**, IFNβ and IFNγ protein levels within the TME of *Adar1*-null and control B16 tumours; $n = 4$ for each condition. **a**, **b**, **f**, Two-sided Student's *t*-test; $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$.

© 2019 Springer Nature Limited. All rights reserved.

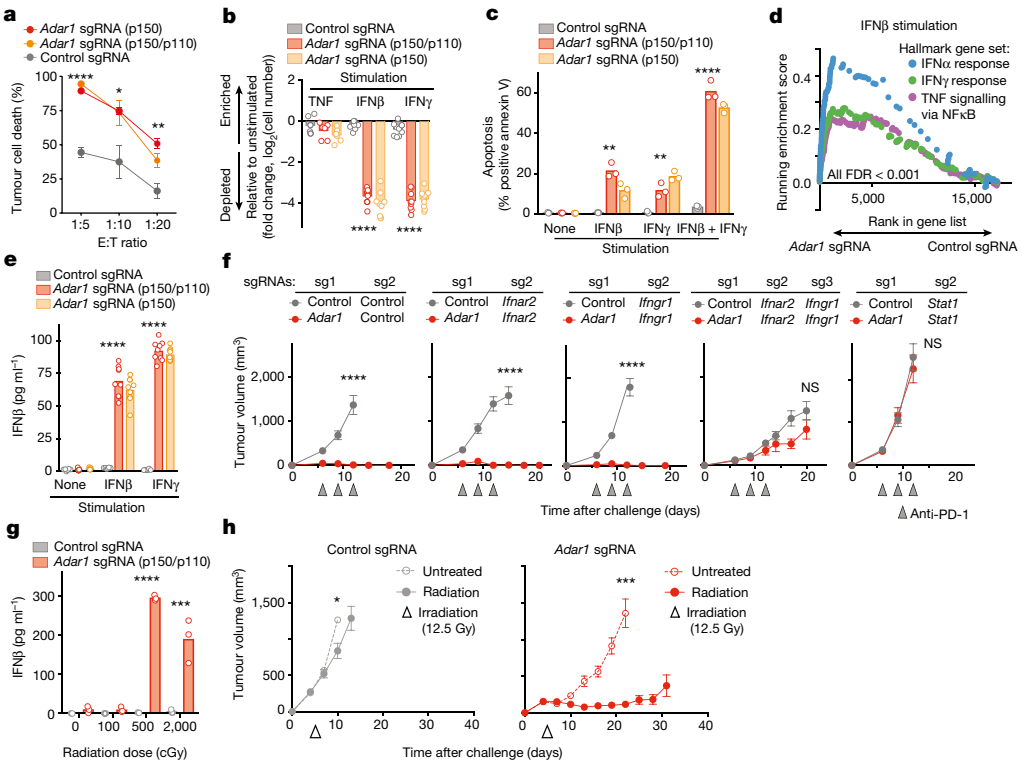


Fig. 3 | Exogenous IFN is required to trigger anti-tumour immunity in *Adar1*-null tumours. a, T cell-dependent killing of ovalbumin-expressing *Adar1* p150/p110-null (red), p150-null (orange) and control (grey) B16 tumour cells by OT-I transgenic T cells specific for an ovalbumin-derived peptide in the context of MHC-I at decreasing E:T ratios (1:20, 1:10 and 1:5). Data shown are representative of two independent experiments with $n = 3$ replicates for each condition; mean ± s.e.m. **b**, Relative numbers of control, *Adar1* p150/p110-null, and *Adar1* p150-null B16 tumour cells stimulated with cytokines indicated compared with unstimulated conditions ($n = 3$ for each condition; data are representative of three independent experiments). **c**, Annexin V staining in control, *Adar1* p110/p150-null and *Adar1* p150-null tumours following stimulation with IFNβ, IFNγ or IFNβ + IFNγ ($n = 3$ for each condition; data are representative of three independent experiments). **d**, GSEA of gene signatures in *Adar1*-null compared with control B16 tumour cells after in vitro culture with IFNβ stimulation. $n = 3$ for each condition; FDR calculated using GSEA. **e**, Enzyme-linked immunosorbent assay (ELISA) for IFNβ in supernatant from control, *Adar1* p150/p110-null and *Adar1* p150-null B16 tumour cells after in vitro culture in unstimulated, IFNβ or IFNγ conditions ($n = 3$ for each condition; data are representative of three independent experiments). **f**, Tumour volume following treatment with anti-PD-1 in vivo in genetic epistasis tumour models that lack ADAR1 and components of IFN sensing pathways ($n = 5$ mice in each group, data representative of two independent experiments). **g**, ELISA for IFNβ in supernatant from control and *Adar1* p150/p110-null B16 tumour cells following irradiation in vitro ($n = 3$ for each condition; data representative of two independent experiments). **h**, Tumour volume of control (grey) and *Adar1*-null (red) B16 tumours following therapeutic irradiation in vivo. $n = 10$ mice in each group; data are representative of two independent experiments. **b**, **c**, **e**, **g**, Bars represent mean; **f**, mean ± s.e.m. **a**–**c**, **e**, **f**, **h**, Two-sided Student's *t*-test; $*P < 0.05$; $**P < 0.01$; $***P < 0.001$, $****P < 0.0001$; NS, not significant.

effector:target (E:T) ratios of 1:5, 1:10 and 1:20, respectively, Student's *t*-test; Fig. 3a, Extended Data Fig. 5a). Enhanced T cell killing of *Adar1*-null tumours could result either from improved T cell cytotoxicity or from an increased sensitivity to secreted effector cytokines such as IFNγ. We tested the latter possibility by stimulating *Adar1*-null and control tumour cells with IFNβ, IFNγ, or TNF (also known as TNFα) and comparing their growth with that of unstimulated cells. Compared with control tumour cells, *Adar1*-null cells showed significant inhibition of viability ($P < 0.0001$, Student's *t*-test) and increased apoptosis ($P < 0.01$, Student's *t*-test) when stimulated with IFNβ or IFNγ, but not TNF (Fig. 3b, c, Extended Data Fig. 5b, c). Similar results were seen in CT26 and *Braf*/*Pten* tumours (Extended Data Fig. 5d, e). Thus, the sensing of either type I or type II IFNs is sufficient to cause growth arrest and apoptosis in *Adar1*-deficient tumour cells.

Adar1-null B16 cells showed significant upregulation of gene signatures of response to IFNα, IFNγ and TNF relative to control cells when cultured with IFNβ or IFNγ (Fig. 3d, Extended Data Fig. 5f, all FDR < 0.001). Cytokine and chemokine genes such as *Ifnb1*, *Il6*, *Ccl5*, *Cxcl9* and *Cxcl10* were upregulated in *Adar1*-null cells following IFN stimulation (Extended Data Fig. 5g). Consistent with this, we found that *Adar1*-null tumours secreted IFNβ following stimulation with IFNβ or IFNγ, whereas control cells or unstimulated *Adar1*-null cells did not (Fig. 3e, Extended Data Fig. 5h). Re-expression of *Adar1* p150 restored IFN-induced growth inhibition and IFNβ secretion to control levels, demonstrating that these phenotypes were unlikely to be off-target effects of gene editing (Extended Data Fig. 6a, b). Thus, *Adar1*-null tumour cells increase expression of antiviral cytokines and chemokines following exposure to IFN.

Sensitivity to immunotherapy requires IFN sensing

We tested whether *Adar1*-null tumours required IFN sensing for enhanced sensitivity to immunotherapy in vivo. We generated *Adar1*-null tumour cell lines that also lacked *Ifnar2*, *Ifngr1* or *Stat1* (double knockout (DKO) cell lines), and triple knockout (TKO) *Adar1*-null tumour cell lines in which both *Ifnar2* and *Ifngr1* were deleted (Extended Data Fig. 6c). The in vitro growth arrest and IFNβ secretion phenotypes seen in *Adar1*-null tumours following stimulation with IFNβ and IFNγ stimulation were abolished by concomitant deletion of *Ifnar2* and *Ifngr1*, respectively (Extended Data Fig. 6d). Similarly, deletion of *Stat1* suppressed both in vitro phenotypes following stimulation with either IFN (Extended Data Fig. 6d). In vivo, genetic deletion of either *Ifnar2* or *Ifngr1* was not sufficient to suppress the sensitivity of *Adar1*-null tumours to PD-1 checkpoint blockade (Fig. 3f). However, concomitant deletion of both *Ifnar2* and *Ifngr1*, or deletion of *Stat1*, abolished the sensitivity of *Adar1*-null tumours to immunotherapy

© 2019 Springer Nature Limited. All rights reserved.

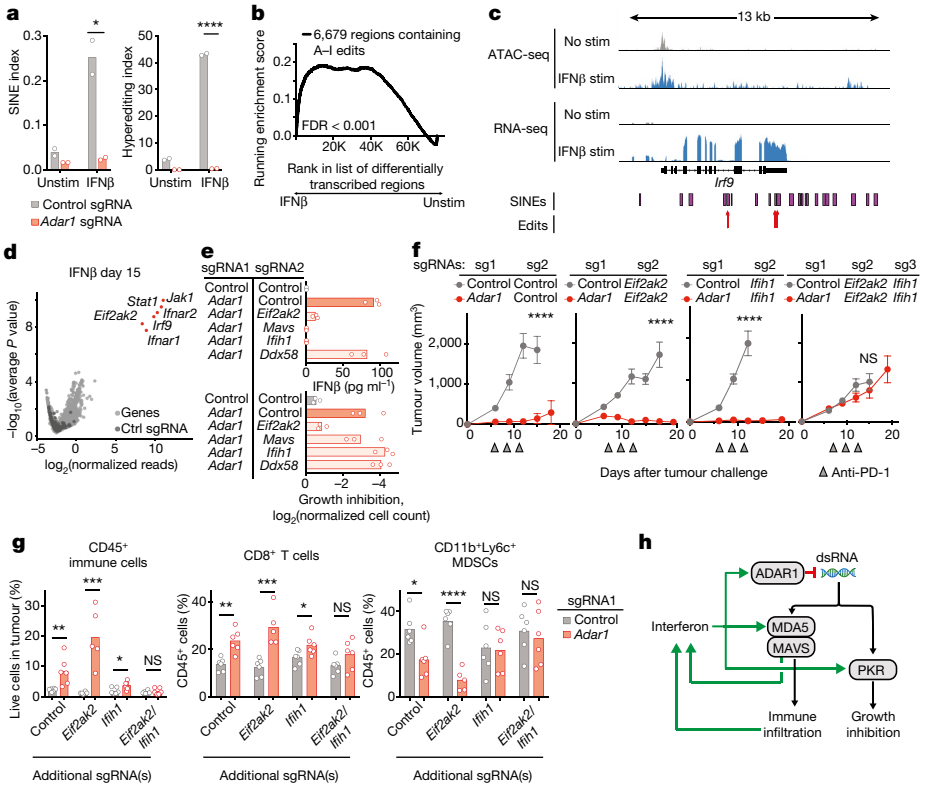


Fig. 4 | dsRNA sensing triggers distinct mechanisms of anti-tumour immunity through MDA5 and PKR in *Adar1*-null tumours. **a**, Quantification of A-to-I editing in SINEs (left) and hyper-editing (right) in control and *Adar1*-null B16 tumours with and without stimulation with IFNβ ($n = 3$ for each condition). **b**, Enrichment in the expression of RNA transcripts containing A-to-I editing sites following stimulation with IFNβ relative to the unstimulated state in control B16 tumour cells. $n = 3$ for each condition; FDR calculated using GSEA. **c**, Assay for transposase accessible chromatin with high-throughput sequencing (ATAC-seq) and RNA-seq at the *Irf9* locus indicating positions of SINEs and A-to-I edits of RNA in IFNβ-stimulated or unstimulated control B16 cells. **d**, Volcano plot depicting the relative depletion and enrichment of sgRNAs targeting 20,146 genes in a Cas9$^+$ *Adar1*-null B16 tumour cell line following stimulation with IFNβ in vitro. *P* values were derived using STARS v1.3. **e**, IFNβ secretion (top) and IFNγ-mediated growth inhibition (bottom) in *Adar1*-null (red) and control (grey) B16 cells with additional deletion of *Eif2ak2* (PKR), *Mavs* (MAVS), *Ifih1* (MDA5), or *Ddx58* (RIG-I) ($n = 3$ for each condition; data are representative of two independent experiments). **f**, Tumour volume following anti-PD-1 treatment of B16 tumours with the genetic perturbations indicated ($n = 5$ mice per group; data are representative of two independent experiments). **g**, Flow cytometry of immune populations from untreated control and *Adar1*-null B16 tumours with the genetic perturbations indicated ($n = 5$ mice per group). **h**, Schema of the genetic dependencies of enhanced immune infiltration and enhanced susceptibility to immune cells in *Adar1*-null tumours. **a**, **e**, **g**, Bars represent mean; **f**, mean ± s.e.m. **a**, **f**, **g**, Two-sided Student's *t*-test; $*P < 0.05$; $**P < 0.01$; $***P < 0.001$; $****P < 0.0001$; NS, not significant.

(Fig. 3f). Thus, *Adar1*-null tumours have an obligate requirement for IFN to mediate their sensitivity to immunotherapy.

ADAR1 loss sensitizes tumours to irradiation

We reasoned that loss of ADAR1 may sensitize tumours to other cancer therapies that are known to induce IFN production in the TME, such as radiation therapy[10–12] or toll-like receptor agonists[13–15]. In vitro, *Adar1*-null tumour cells produced significantly more IFNβ than did control tumours after irradiation (Fig. 3g, Extended Data Fig. 6e) and were significantly less viable after irradiation (Extended Data Fig. 6e; $P < 0.001$, Student's *t*-test). Radiation (12.5 Gy) or topical therapy with imiquimod significantly slowed tumour growth and enhanced survival in animals bearing *Adar1*-null tumours, but had a minimal effect on control B16 tumours (Fig. 3h, Extended Data Fig. 6d, f, g; $P < 0.0001$ for survival in both cases, log-rank test). Thus, loss of ADAR1 increases the efficacy of therapies that can elicit the production of IFN in the TME.

ADAR1-edited RNAs are preferentially induced by IFN

We reasoned that the requirement for IFN to trigger the observed response in *Adar1*-null cells might be explained by the IFN-mediated upregulation of RNA species that are normally edited by ADAR1 and that can serve as ligands for dsRNA sensors. We found a larger number of A-to-I editing events in small interspersed nuclear elements (SINEs) and a greater abundance of RNA hyperediting following IFN stimulation in control cells compared with *Adar1*-null cells (Fig. 4a; $P < 0.05$ and $P < 0.0001$, respectively, Student's *t*-test). SINEs containing known ADAR1 edit sites were enriched around genes, in particular 3′ UTRs, compared with the genomic distribution of all SINEs[16,17] (Extended Data Fig. 7a; $P < 1 \times 10^{-30}$, hypergeometric test).

We found that RNA species with evidence of A-to-I edits were significantly upregulated by stimulation with IFNβ (Fig. 4b, c, Extended Data Fig. 7b; FDR < 0.001). Consistent with this, we found a highly significant association between edited sites adjacent to IFN-inducible regions of accessible chromatin (data not shown; $P = 0.0035$, one-sided binomial test). Thus, IFNs increase the transcription of RNA species that are normally edited by ADAR1 and that may serve as ligands for dsRNA sensors. Together with increased abundance of the RNA sensors themselves (Extended Data Fig. 7c), this may explain the requirement for IFN to reveal the vulnerability of *Adar1*-null tumour cells.

PKR and MDA5 mediate distinct phenotypes

We conducted a genome-wide screen using CRISPR to identify genes required for the IFN-induced growth arrest phenotype of *Adar1*-null tumour cells. Following culture with IFNβ, we found significant enrichment of sgRNAs targeting genes required for the sensing of type I IFN (*Ifnar1*, *Ifnar2*, *Jak1*, *Stat1*, and *Irf9*; FDR < 0.0002, Fig. 4d) in the surviving cells. We also found marked enrichment of sgRNAs targeting *Eif2ak2*, the gene that encodes protein kinase R (PKR).

© 2019 Springer Nature Limited. All rights reserved.

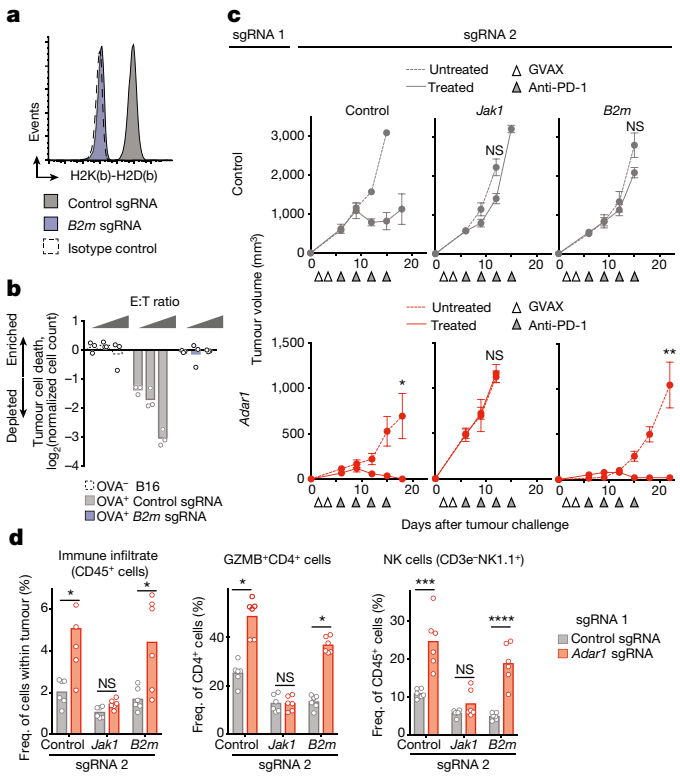


Fig. 5 | Loss of ADAR1 overcomes resistance to immunotherapy mediated by loss of antigen presentation. a, MHC I expression in control (grey) and *B2m*-null (purple) B16 tumour cell lines relative to isotype control (dashed line). Data are representative of two independent experiments. **b**, T cell killing assay with non-ovalbumin-expressing control (dashed line), ovalbumin-expressing control (grey) and ovalbumin-expressing *B2m*-null (purple) B16 cells ($n = 3$ for effector target ratios 1:20, 1:10 and 1:5 for each condition; data shown are representative of two independent experiments). **c**, Tumour volume following anti-PD-1 and GVAX treatment of B16 tumours with the genetic perturbations indicated. Data (mean ± s.e.m.) are representative of two independent experiments with $n = 5$ mice in each group. **d**, Flow cytometry of immune populations following anti-PD-1 and GVAX treatment of B16 tumours with the genetic perturbations indicated ($n = 5$ mice per group). Bars represent mean. **c**, **d**, Two-sided Student's *t*-test; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$; NS, not significant.

In a second screen we used IFNγ to mediate growth arrest and found enrichment of sgRNAs targeting *Jak1*, *Stat1*, and *Efi2ak2* but not those targeting *Ifngr1* or *Ifngr2*. We did, however, observe enrichment of sgRNAs targeting *Ifnar1* and *Ifnar2* (FDR < 0.0007, Extended Data Fig. 7d), suggesting that IFNγ stimulation elicits type I IFN secretion from *Adar1*-deficient cells, which in turn mediates growth arrest via PKR. Thus, sensing of dsRNA by PKR is the major mechanism that underlies the growth arrest phenotype of *Adar1*-null tumour cells following IFN stimulation.

To identify the genes required for the secretion of IFNβ by *Adar1*-null cells, we generated DKO *Adar1*-null cell lines that also lacked either *Ifih1* (MDA5), *Ddx58* (RIG-I), *Mavs* (MAVS), or *Eif2ak2* (PKR) and tested them for suppression of the growth arrest and IFNβ secretion phenotypes (Fig. 4e, Extended Data Fig. 7e). Loss of PKR abolished the IFN-induced growth arrest phenotype. Loss of other dsRNA sensors had no effect on IFN-mediated growth arrest (Fig. 4e). Deletion of MDA5 or MAVS completely suppressed secretion of IFNβ, which was also reduced in *Adar1*-null tumour cells that lacked PKR (Fig. 4e). This suggests that MDA5 and MAVS do not suppress IFN-mediated growth arrest in *Adar1*-null cells, but are required for IFNβ secretion.

We next tested which dsRNA sensor was responsible for the enhanced response of *Adar1*-null tumours to immunotherapy in vivo. DKO *Adar1*-null cell lines that also lacked either PKR or MDA5 were as sensitive to immunotherapy as *Adar1*-null single knockout tumours (Fig. 4f). However, TKO tumour cells that lacked ADAR1, PKR and MDA5 no longer showed enhanced sensitivity to immunotherapy (Fig. 4f). Thus, dsRNA sensing through either PKR or MDA5 is sufficient to confer the enhanced response to checkpoint blockade of *Adar1*-null tumours, but loss of both pathways abrogates this increased sensitivity.

We next tested which dsRNA sensor was required for the inflammation observed in *Adar1*-null tumours in vivo. *Adar1*-null tumours that lacked PKR showed similar or greater inflammation of the TME compared with *Adar1*-null tumours and greater inflammation than control tumors (Fig. 4g). By contrast, *Adar1*-null tumours lacking MDA5 showed only a minor increase in inflammation relative to control tumours, and *Adar1*-null tumours lacking both PKR and MDA5 showed no increase in immune infiltration (Fig. 4g). Similarly, whereas both IFNβ and IFNγ were significantly increased in the microenvironments of *Adar1*-null tumours with or without PKR ($P < 0.001$ for both following IFNβ stimulation and $P < 0.0001$ for both following IFNγ stimulation), we detected no increase in IFN in *Adar1*-null tumours lacking MDA5 or in *Adar1*-null tumours lacking both PKR and MDA5 (Extended Data Fig. 7f). Thus dsRNA sensing by MDA5 is required for the enhanced inflammation and immune infiltration in *Adar1*-null tumours, but activation of PKR is not. MDA5 and PKR therefore mediate distinct mechanisms of increased susceptibility to anti-tumour immunity in *Adar1*-null tumours (Fig. 4h). Consistent with these results, levels of hyperediting in human tumours were inversely correlated with immune infiltration and inflammatory response (Extended Data Fig. 8a, b, Supplementary Text II).

Loss of ADAR1 overcomes resistance to immunotherapy

We tested whether deletion of *Adar1* was sufficient to overcome common mechanisms of resistance to checkpoint blockade. To generate mouse models of immunotherapy resistance, we deleted *B2m* in B16 tumours and confirmed that this abolished recognition of tumour cells by $CD8^+$ T cells (Fig. 5a, b, Extended Data Fig. 9a, b) and rendered them completely resistant to immunotherapy in vivo (Fig. 5c, top; Extended Data Fig. 10a).

We next generated cell lines lacking both *Adar1* and *B2m* and compared their sensitivity in vivo to immunotherapy with granulocyte–macrophage colony-stimulating factor (GM-CSF)-secreting whole tumour cell vaccine (GVAX) and PD-1 blockade with those of cell lines lacking only *B2m*. Loss of ADAR1 restored sensitivity to immunotherapy and resulted in elimination of many of the *B2m*-null, resistant tumours (Fig. 5c, Extended Data Fig. 10a). Similar results were found for other resistance mutations: *H2-K1*, *Nlrc5* and *Jak2* (Extended Data Figs. 9b, 10b, c). By contrast, tumour cells in which resistance had been engineered by deletion of *Jak1* were not re-sensitized by loss of ADAR1 (Fig. 5c), underscoring the essential role of IFN sensing in the sensitivity of *Adar1*-null tumours to immunotherapy.

Loss of ADAR1 in *Jak1*-null tumours did not change the composition of immune cells in the tumours. However, loss of ADAR1 in *B2m*-null tumours was associated with a significant increase in immune cells, including non-MHC I-restricted cytotoxic cells such as γδ T cells, granzyme B^+ $CD4^+$ T cells and NK cells (Fig. 5d, Extended Data Fig. 10d, $P < 0.05$, $P < 0.001$, $P < 0.001$, $P < 0.0001$, respectively, Student's *t*-test), and a significant decrease in suppressive myeloid cells (Extended Data Fig. 10d, $P < 0.01$, Student's *t*-test). Thus, deletion of *Adar1* overcomes several common mechanisms of acquired resistance to immunotherapy and causes inflammatory repolarization of the TME even in the absence of $CD8^+$ T cell recognition of MHC I on tumours.

Discussion

We have shown that loss of function of ADAR1 in tumour cells removes a checkpoint that normally restrains sensing of IFN-inducible dsRNA, leading to enhanced tumour inflammation and heightened IFN sensitivity mediated by MDA5 and PKR, respectively. This dual mechanism

© 2019 Springer Nature Limited. All rights reserved.

increases the response to immune checkpoint blockade and overcomes resistance to immunotherapy.

Because tumours without immune infiltration often fail to respond to checkpoint blockade, strategies to inflame the TME are a high therapeutic priority. These include increased delivery or epigenetic derepression of innate ligands[10,18–22]. However, our data suggest that cancer cells already contain, or can be induced to express, sufficient quantities of immunostimulatory nucleic acids to increase tumour immunity and overcome resistance to checkpoint blockade, if the mechanisms that limit their detection can be overcome.

Effective immunotherapy with checkpoint blockade is assumed to require recognition of tumour cells by cytotoxic $CD8^+$ T cells[23–25], but our results show that loss of function of ADAR1 restores sensitivity to immunotherapy in tumours with a *B2m* deletion. This suggests that recognition of tumours by $CD8^+$ T cells is not an obligate part of an effective immune response against cancer cells. Rather, loss or lack of antigen presentation can be overcome if sufficient inflammation can be elicited in an IFN-sensitive tumour. This finding suggests a strategy for effective immunotherapy even in the absence of a tumour-specific endogenous $CD8^+$ T-cell response.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0768-9.

Received: 20 February 2018; Accepted: 8 November 2018;
Published online 17 December 2018.

1. Zaretsky, J. M. et al. Mutations associated with acquired resistance to PD-1 blockade in melanoma. *N. Engl. J. Med.* **375**, 819–829 (2016).
2. Gao, J. et al. Loss of IFN-γ pathway genes in tumor cells as a mechanism of resistance to anti-CTLA-4 therapy. *Cell* **167**, 397–404 (2016).
3. Yoshihama, S. et al. NLRC5/MHC class I transactivator is a target for immune evasion in cancer. *Proc. Natl Acad. Sci. USA* **113**, 5999–6004 (2016).
4. Sade-Feldman, M. et al. Resistance to checkpoint blockade therapy through inactivation of antigen presentation. *Nat. Commun.* **8**, 1136 (2017).
5. Manguso, R. T. et al. *In vivo* CRISPR screening identifies Ptpn2 as a cancer immunotherapy target. *Nature* **547**, 413–418 (2017).
6. Mannion, N. M. et al. The RNA-editing enzyme ADAR1 controls innate immune responses to RNA. *Cell Reports* **9**, 1482–1494 (2014).
7. Liddicoat, B. J., Chalk, A. M. & Walkley, C. R. ADAR1, inosine and the immune sensing system: distinguishing self from non-self. *Wiley Interdiscip. Rev. RNA* **7**, 157–172 (2016).
8. Ahmad, S. et al. Breaching self-tolerance to Alu duplex RNA underlies MDA5-mediated inflammation. *Cell* **172**, 797–810.e13 (2018).
9. Chung, H. et al. Human ADAR1 prevents endogenous RNA from triggering translational shutdown. *Cell* **172**, P811–P824 (2018).
10. Deng, L. et al. STING-dependent cytosolic DNA sensing promotes radiation-induced type I interferon-dependent antitumor immunity in immunogenic tumors. *Immunity* **41**, 843–852 (2014).
11. Lim, J. Y. H., Gerber, S. A., Murphy, S. P. & Lord, E. M. Type I interferons induced by radiation therapy mediate recruitment and effector function of $CD8^+$ T cells. *Cancer Immunol. Immunother.* **63**, 259–271 (2014).
12. Twyman-Saint Victor, C. et al. Radiation and dual checkpoint blockade activate non-redundant immune mechanisms in cancer. *Nature* **520**, 373–377 (2015).
13. Wenzel, J., Uerlich, M., Haller, O., Bieber, T. & Tueting, T. Enhanced type I interferon signaling and recruitment of chemokine receptor CXCR3-expressing lymphocytes into the skin following treatment with the TLR7-agonist imiquimod. *J. Cutan. Pathol.* **32**, 257–262 (2005).
14. Sainathan, S. K. et al. Toll-like receptor-7 ligand Imiquimod induces type I interferon and antimicrobial peptides to ameliorate dextran sodium sulfate-induced acute colitis. *Inflamm. Bowel Dis.* **18**, 955–967 (2012).
15. Cho, J. H. et al. The TLR7 agonist imiquimod induces anti-cancer effects via autophagic cell death and enhances anti-tumoral and systemic immunity during radiotherapy for melanoma. *Oncotarget* **8**, 24932–24948 (2017).
16. Patterson, J. B. & Samuel, C. E. Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. *Mol. Cell. Biol.* **15**, 5376–5388 (1995).
17. George, C. X. & Samuel, C. E. Human RNA-specific adenosine deaminase ADAR1 transcripts possess alternative exon 1 structures that initiate from different promoters, one constitutively active and the other interferon inducible. *Proc. Natl Acad. Sci. USA* **96**, 4621–4626 (1999).
18. Heidegger, S. et al. The RIG-I agonist 3pRNA synergizes with checkpoint blockade in cancer immunotherapy. *Blood* **126**, 3436 (2015).
19. Goel, S. et al. CDK4/6 inhibition triggers anti-tumour immunity. *Nature* **548**, 471–475 (2017).
20. Chiappinelli, K. B. et al. Inhibiting DNA methylation causes an interferon response in cancer via dsRNA including endogenous retroviruses. *Cell* **162**, 974–986 (2015).
21. Li, H. et al. Immune regulation by low doses of the DNA methyltransferase inhibitor 5-azacitidine in common human epithelial cancers. *Oncotarget* **5**, 587–598 (2014).
22. Ribas, A. et al. Oncolytic virotherapy promotes intratumoral T cell infiltration and improves anti-PD-1 immunotherapy. *Cell* **170**, 1109–1119 (2017).
23. Juneja, V. R. et al. PD-L1 on tumor cells is sufficient for immune evasion in immunogenic tumors and inhibits CD8 T cell cytotoxicity. *J. Exp. Med.* **214**, 895–904 (2017).
24. McGranahan, N. et al. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* **171**, 1259–1271 (2017).
25. Tumeh, P. C. et al. PD-1 blockade induces responses by inhibiting adaptive immune resistance. *Nature* **515**, 568–571 (2014).

Acknowledgements We thank all members of the Haining, Hur, Levanon and Meyerson laboratories for their input and discussions regarding this project.

Reviewer information *Nature* thanks C. Walkley and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions W.N.H., J.J.I., R.T.M. and C.K.C. contributed to the study design. J.J.I., C.K.C., D.E.C., R.T.M., A.A., I.C.K., H.W.P., M.D.Z., S.K.L.-R., E.J.R. and A.H.L. performed validation experiments in cell lines and with live animals. C.K.C., A.I.-V. and R.T.M. carried out ADAR1 re-expression/rescue experiments. C.K.C., J.J.I., R.T.M. and M.D.Z. conducted epistasis experiments. A.P. and F.D.B. performed cell death experiments. J.J.I. and D.K. designed and performed radiation experiments. J.J.I., B.C.M. and G.K.G. conducted and analysed immunohistochemical experiments. J.J.I., J.D., A.I.-V. and B.C.M. conducted flow cytometry experiments, which B.C.M. and J.D. analysed. J.J.I. and B.C.M. prepared samples for single-cell RNA-seq, which K.B. analysed. J.J.I. and K.B.Y. prepared samples for RNA-seq analysis. K.B.Y., I.B., E.Y.L., A.P., P.P.D., K.B. and D.R.S. performed editing, RNA-seq and ATAC-seq analysis. J.G.D. performed screening processing and analysis. J.J.I., N.B.C., K.B. and A.P. undertook analysis of human tumours. J.J.I., R.T.M. and W.N.H. wrote the manuscript.

Competing interests This work was supported in part by funding from Calico Life Sciences, LLC. J.J.I., R.T.M. and W.N.H. are authors of a patent application related to ADAR. W.N.H. consults for and has equity in Tango Therapeutics.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0768-9.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0768-9.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to W.N.H.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

Creation of CRISPR-edited tumour cell lines. *Adar1* was deleted in Cas9-expressing B16, *Braf/Pten* and MC38 mouse tumour cell lines for validation experiments using a lentiviral delivery system (pXPR_BRD024, Addgene) to express sgRNAs and puromycin selection as previously described[5]. For further validation experiments, confirmatory epistasis and re-expression/rescue experiments, *Adar1* was deleted in B16 and CT26 cells using transient transfection of a Cas9-sgRNA plasmid (pX459, Addgene) with the Turbofect transfection reagent (Thermo Fisher Scientific, R0531) or lipofectamine transfection reagent (Thermo Fisher Scientific, L3000015), respectively, and puromycin selection. For epistasis experiments, Cas9 was expressed using the pLX311 backbone, transient transfection was used to introduce the first guide(s), and the final epistasis guides were expressed using the pXPR_BRD024 lentiviral expression system. For in vitro re-expression/rescue experiments, *Adar1* was synthesized from the mm10 consensus sequence and either ADAR1 or an irrelevant control protein (CD19) was expressed using the pLX311 backbone used in prior work to express Cas9[5]. Cell lines were tested every 3–6 months for mycoplasma contamination.

Animal treatment and tumour challenges. The designs of animal studies and procedures were approved by the Dana Farber Cancer Institute IACUC and the Broad Institute IACUC committees. Ethical compliance with IACUC protocols and institute standards was maintained. Specific pathogen-free facilities at the Dana Farber and Broad Institutes were used for the storage and care of all mice. Six-week-old wild-type female C57BL/6J mice were obtained from Jackson laboratories. A colony of B6.129S2-Tcra^{tm1Mom}/J (Tcra) T cell-deficient mice were bred on site at the Dana Farber Institute. A colony of NOD.Cg-*Prkdc*^{scid} *Il2rg*^{tm1Wjl}/SzJ (NSG) mice were bred on site at the Broad Institute. Mice were age-matched to be 6–12 weeks old at the time of tumour inoculation. For tumour challenges, 2.0×10^6 tumour B16, *Braf/Pten* or MC38 cells were resuspended in Hanks balanced salt solution (Gibco), mixed 1:1 by volume with matrigel (Corning) and subcutaneously injected into the right flank on day 0. CT26 cells (1.0×10^6) were resuspended in Hanks balanced salt solution and injected as described above. Each tumour injected contained only a single sgRNA targeting each indicated gene. Where indicated, mice were vaccinated with 1.0×10^6 GM-CSF-secreting B16 (GVAX) cells (kindly provided by G. Dranoff) that had been irradiated with 35 Gy on days 1 and 4 to elicit an anti-tumour immune response. For non-resistance validation experiments, mice were treated with 100 μg of rat monoclonal anti-PD1 antibodies (Bio X Cell, clone: 29F.1A12) on days 6, 9 and 12 (for B16 and CT26) or day 9 (for MC38) via intraperitoneal injection. Rat IgG2a isotype control was used in control mice corresponding to the anti-PD1 treatment group for B16 experiments shown in Fig. 1. For resistance experiments, mice were treated with 200 μg of rat anti-PD1 antibodies on days 6, 9, 12 and 15. Each tumour was measured every 3–4 days beginning on day 6 after challenge until either the survival endpoint was reached or no palpable tumour remained. Measurements were assessed manually by assessing the longest dimension (length) and the longest perpendicular dimension (width). Tumour volume was estimated with the formula: $(L \times W^2)/2$. CO_2 inhalation was used to euthanize mice. For irradiation experiments, the Dana Farber Small Animal Radiation Research Platform was used[26]. In brief, mice were anesthetized via isoflurane inhalation for the duration of each treatment. For each treatment, tumours were visualized using cone beam computed tomography (CT) using 60 kVp and 0.8 mA photons. Tumours were treated using a 10 × 10-mm square shaped collimator selected to give 0.25–0.5-cm margins around gross tumour, using 220 kVp and 13 mA photons given with a lateral en face field prescribed to a depth of 5 mm. The Small Animal Radiation Research Platform was calibrated and maintained as previously described[26]. For imiquimod experiments, 5% imiquimod cream was obtained through the Dana-Farber Cancer Institute animal facility. A thin film of imiquimod cream was applied to the skin overlying tumours every three days following tumour inoculation until tumour outgrowth or disappearance. No statistical methods were used to predetermine sample size. For all experiments, at least five mice were included in each group, based upon prior knowledge of the variability of experiments with immune checkpoint blockade. Animals were randomized before treatment and no blinding was performed.

Tumour size match experiment. We implanted 2×10^6 *Adar1*-null (*Adar1* sgRNA 2) cells in matrigel or 1×10^6 control B16 tumour cells in HBSS into the right flank of 6- to 7-week-old C57BL/6 female mice. Mice were treated with 5 mg/kg anti-PD-1 antibody on days 6, 9 and 12 as described above.

Immunohistochemistry. Immunohistochemical (IHC) staining was performed at the Dana-Farber/Harvard Cancer Center Specialized Histopathology Core using a Leica Bond automated staining platform with anti-CD3 (Abcam, clone ab16669; 1:150 dilution) and anti-CD8 (eBio, clone 14-0808; 1:100 dilution) antibodies. Slides were visualized using Aperio software. $CD3^+$ and $CD8^+$ cells that stained with strong membranous positivity were enumerated in five separate areas at 20× magnification in a blinded fashion by G.K.G. for each slide.

Analysis of tumour-infiltrating lymphocytes by flow cytometry. Control guide or *Adar1*-null tumour cells (*Adar1* sgRNA 2; 2×10^6) were implanted in matrigel into BL6 female mice at 5–8 weeks of age. On day 14 following implantation, tumours were dissected from the surrounding fascia, weighed, mechanically minced, and treated with collagenase P (2 mg/ml, Sigma) and DNase I (50 μg/ml, Sigma) for 10 min at 37 °C. Cells were passed through a 70-micrometre filter to remove clumps, diluted in medium, and a small aliquot taken directly for flow cytometry. Cell surface staining was performed with the indicated antibodies before fixation and permeabilization of the cells (Intracellular Fixation & Permeabilization Buffer Set, eBiosciences) for intracellular staining. Sphero AccuCount Fluorescent Particles (Spherotech) were added to each tube to allow cell counting before analysis on a LSR II flow cytometer (BD Biosciences). All analysis was done with FlowJo software v 10.4.2 (FlowJo). Cell counts were determined by normalizing cell numbers to beads recorded, divided by the amount of tumour aliquot taken and the mass of the tumour. See Supplementary Information Table 1 for a full list of antibodies used.

Analysis of tumour-infiltrating lymphocytes by single cell RNA-seq. *Adar1*-null (sgRNA2) or control tumour cells (2×10^6) were implanted in matrigel into the right flank of C57BL/6 female mice. On day 14, tumours were dissected from the surrounding fascia, mechanically minced, and treated with collagenase P (2 mg/ml, Sigma) and DNase I (50 μg/ml, Sigma) for 10 min at 37 °C. Tumour-infiltrating leukocytes were enriched using an Optiprep (Sigma) density gradient followed by $CD45^+$ MACS positive selection (Miltenyi). B16 tumour cells grown in culture were added to each sample at a 5% ratio as a spike-in control for assessing sample-to-sample variability. Cells were counted and loaded onto the 10x device (10x Genomics). Samples were processed per the manufacturer's protocol and sequenced on an Illumina NextSeq sequencer. Sample demultiplexing, barcode processing, alignment, filtering, UMI counting, and aggregation of sequencing runs were performed using the Cell Ranger analysis pipeline (v1.2). Downstream analyses were performed in R using the Seurat package[27].

For each cell, two quality control metrics were calculated: (1) the total number of genes detected and (2) the proportion of UMIs contributed by mitochondrially encoded transcripts. Cells in which fewer than 200 genes were detected and in which mitochondrially encoded transcripts constituted more than 10% of the total library were excluded from downstream analysis. Genes detected in fewer than three cells across the data set were also excluded, yielding a preliminary expression matrix of 8,834 cells (comprised of both infiltrating immune cells and spiked-in tumour cells) by 17,190 genes. To assess technical variability between samples, an initial *t*-SNE projection was generated using all 8,834 cells (data not shown); co-clustering of spiked-in tumour cells expressing *Pmel* and *Mlana* (transcriptional markers of melanoma) from all four experiments demonstrated minimal sample-to-sample variability. We subsequently removed 1,428 tumour cells from the total expression matrix, leaving only infiltrating immune cells for downstream analysis.

Mean and dispersion values were calculated for each gene across the remaining 7,406 cells, and a subset of 1,494 highly variable genes was selected for principal components analysis (PCA). Following PCA, the first 55 PCs were determined to be significant ($P < 0.01$) using the jackstraw method and *t*-SNE was performed on these significant PCs using default parameters for 1,000 iterations for visualization in two dimensions. Unsupervised clustering using a shared nearest neighbour modularity optimization based algorithm (resolution parameter 0.8) identified 15 distinct clusters[28]. See Supplementary Information Table 2 for cell cluster and barcode identification for each cell. For classification of immune cell populations, differential expression analysis was performed between each cluster and all other cells using a Wilcoxon rank sum test. Top differential expression results for each cluster were cross-referenced with canonical markers for a comprehensive range of immune cell populations, yielding a consensus panel of transcriptional markers for each of the 15 clusters (Extended Data Fig. 4a, b, Supplementary Information Table 3).

For preranked GSEA, differential expression analysis was performed between all infiltrating immune cells from *Adar1*-null tumours and control tumours using a Wilcoxon rank sum test, and a ranking metric was calculated for each gene as $R = -\log_{10}(q)$, where q is the FDR-adjusted P value (Supplementary Information Table 4). Preranked GSEA was performed using a curated collection of gene sets consisting of sets from the Hallmark and Gene Ontology collections in the MSigDB database[29]. Single-cell signature scoring using FastProject was also performed using this curated collection[30].

RNA-seq analysis of tumour cells. *Adar1*-null or control sgRNA-transfected B16 cells were stimulated with IFNβ (1,000 U/ml, PBL) for 36 h. RNA was extracted from cell pellets using the Qiagen RNeasy Mini kit according to the manufacturer's instructions. First-strand Illumina-barcoded libraries were generated using the NEB RNA Ultra Directional kit according to the manufacturer's instructions, using ribosomal RNA depletion and including a 12-cycle PCR enrichment. Libraries were sequenced on an Illumina NextSeq 500 instrument using paired-end 37-bp reads. Data were trimmed for quality using the Trimmomatic pipeline with the following parameters: LEADING:15 TRAILING:15 SLIDINGWINDOW:4:15

© 2019 Springer Nature Limited. All rights reserved.

MINLEN:16. Data were aligned to mouse reference genome mm10 using Bowtie2. HTSeq was used to map aligned reads to genes and to generate a gene count matrix. Normalized counts and differential expression analysis was performed using the DESeq2 R package (Supplementary Information Table 5). GSEA was performed as previously described, using the Hallmark gene signature collection[29,31] (Supplementary Information Table 6).

RNA editing analysis of tumour cells. All editing analysis was performed using tumour cell RNA-seq data, which were generated as indicated above. The quality of the sequence reads was confirmed using the FastQC (https://www.bioinformatics.babraham.ac.uk) quality control tool with default parameters. Duplicated reads were removed using prinseq[32]. Next, sequence reads were aligned using the STAR[33] aligner to the mm9 reference genome with parameters that accept only uniquely aligned reads (outFilterMultimapNmax = 1) and limit the number of mismatches to 0.05 of the mapped length (outFilterMismatchNoverLmax = 0.05).

In order to generate the SINE index measurements, a previously published human Alu-specific editing detection algorithm[34] was adjusted to screen three mouse SINE subfamilies: B1, B2 and B4. Similar to the Alu editing index, the SINE editing index is defined as the number of guanosines that were aligned to genomic adenosines that reside in SINE element, divided by the total number of nucleotides within the reads that align to SINE adenosine positions. All edited sites were converted to mm10 genome coordinates for downstream analysis.

Hyper-editing analysis is another global estimate of RNA editing levels[35]. This analysis quantifies heavily edited (hyper-edited) reads, which fail to align to the corresponding genome using standard alignment tools, and are hence traditionally overlooked. In order to align these hyper-edited reads, we transformed all adenosines to guanosines in both the unmapped reads and the reference genome and realigned them, and then transformed back the nucleotides to identify all mismatches. For each sample, the number of hyper-edited reads per million mapped reads is used to quantify the level of hyper-editing. All hyper-edited sites were converted to mm10 genome coordinates for downstream analysis.

Reads from the RNA-seq, which were generated as described above, were then aligned to the mouse reference genome (mm10) using STAR 2.5.4a aligner with the following parameters that accept only uniquely aligned reads (outFilterMultimapNmax 1) and limit the absolute number of mismatches per read to 10 (–outFilterMismatchNmax 10). We marked duplicate reads in the bam files using PICARD. Next, we attempted to capture the entire transcriptional universe by using MACS2 version 2.1.0, setting FDR to 0.05 and default mouse genome size, to call peaks on aligned reads from two biological conditions (two replicates each): untreated and IFNβ-treated control sgRNA-transfected B16 cells.

Transcribed regions were defined by using BEDTools to extend MACS2-called peaks across regions of the genome that were covered by at least 10 reads in any sample. These regions then were filtered for lengths of more than 50 bases. Furthermore, transcribed regions were checked for accuracy by viewing them on the Integrated Genomics Viewer and by calculating genome coverage using BEDTools. BEDTools was used to calculate the average read coverage for each transcribed region. Differential expression analysis between transcribed regions was performed using DESeq2. To assess the enrichment of edit site containing transcribed regions in IFNβ-treated versus untreated control sgRNA-transfected B16 cells, GSEAPreranked was performed between these groups using the ranking statistic $k = (-1 \times \log_{10}(q)) \times (\mathrm{Abs}(f)/f)$, where q is the DESeq2 FDR-adjusted P value and f is the signed fold change of a given transcribed region.

Genomic annotation of all SINE elements and edited sites found within SINE elements was performed using the ChIPseeker R package[36].

ATAC-seq analysis of tumour cells. For generation of ATAC-seq data, B16 tumour cells were grown in vitro with and without IFNβ stimulation for 36 h. Fifty thousand cells per replicate were sorted into PBS with 10% FBS. Pelleted cells were lysed in 50 μl reaction mix (25 μl of 2× TD, 2.5 μl of Tn5 enzyme, 0.25 μl of 2% digitonin, 22.25 μl of nuclease-free water). The reaction was incubated at 37 °C for 30 min with agitation at 300 r.p.m. DNA was purified using a QIAgen MinElute Reaction Cleanup kit and Nextera sequencing primers ligated using PCR amplification. Agencourt AMPure XP bead cleanup (Beckman Coulter/Agencourt) was used post-PCR and library quality was verified using a Tapestation machine. Samples were sequenced on an Illumina NextSeq 550 sequencer using paired-end 37-bp reads.

Raw reads in Fastq files were trimmed for quality and primers were removed using Trimmomatic-0.33 using the following parameters: LEADING:15 TRAILING:15 SLIDINGWINDOW:4:15 MINLEN:36. FastQC reports were generated before and after trimming to assess quality. Trimmed reads were aligned to mm10 with bowtie2.2.4. Aligned bams were sorted for marking duplicates, and reads mapping to the blacklist region were removed. Reads were shifted +4 bp and −5 bp using pysam 0.9.0. Bam files from biological replicates were merged using samtools 1.3 before peak-calling using MACS 2.1.1 at a q-value threshold of 0.001. Consensus peaks were merged to create a single peak universe. Cut sites were extracted from each biological replicate and the number of cuts within each peak region was quantified to generate a raw counts matrix. DESeq2 was used to normalize the counts matrix and perform differential accessibility analysis between all pairwise comparisons. Regions were subsetted into those with increased, decreased or non-differential accessibility after IFN treatment compared with baseline samples. All peaks were extended by 250 bp on each side (total extension of 500 bp) using BEDTools2.2.20 and the number of overlaps with previously defined edits (see above) were quantified. The significance of overlap with edits was determined using a one-directional binomial test with non-differential regions as the background.

TCGA hyperediting analysis. A hyperediting index was obtained from a published study[37] characterizing primary tumour samples from 356 patients with publically available RNA-seq data in the TCGA collection (https://cancergenome.nih.gov/). Gene signature scores for Hallmark gene sets[31] were assigned to these primary tumour samples using single-sample gene set enrichment analysis (ssGSEA) and the GenePattern interface (https://genepattern.broadinstitute.org). CIBERSORT[38] was used to calculate an absolute immune infiltrate score for all primary tumour samples. ESTIMATE[39] was used to independently quantitate immune infiltrate for each primary tumour sample (http://bioinformatics.mdanderson.org/estimate/). For samples without a publically available ESTIMATE score, scores were calculated using the ESTIMATE R package. Pearson correlation tests were performed using R.

In vitro cytokine stimulations and growth inhibition assays. Tumour cells were engineered as noted above and plated in DMEM (B16, *Braf*/*Pten* and MC38) or RPMI (CT26) + 10% FBS containing the indicated combinations of cytokines: IFNβ (1,000 U/ml, PBL), IFNγ (100 ng/ml, Cell Signaling Technologies), TNF (10 ng/ml, PreproTech). For cell growth and viability assays, 10,000 cells were plated in 96-well plates and viable cells were enumerated after 72–96 h using Cell Titer-Glo (Promega, G7570). Growth assays depicted in the main figures were repeated for confirmation as follows: 50,000 cells were plated in 12-well plates and viable cells were counted after 72 h using the Countess automated cell counting system (Thermo Fisher Scientific, C10227).

Cell death assays. Three sets of transfected B16 cells (control sgRNA5, *Adar1* sgRNA1 and *Adar1* sgRNA2) were plated in separate 6-well plates at a concentration of 100,000 cells per well and incubated for 72 h with DMEM + 10% FBS containing one of the following combinations of cytokines: IFNβ, IFNγ, or IFNβ and IFNγ. Cytokine-treated B16 cells, following trypsinization and washes in PBS + 2% FBS, were stained for 20 min on ice using the manufacturer's recommended concentrations of Annexin-V PE and 7-AAD from the PE Annexin V Apoptosis Detection Kit 1(BD Pharmingen) and with Calcein-AM (ThermoFisher Scientific). Staining of cell surface markers was then analysed using an Accuri C6 flow cytometry system. Analysis was carried out using FlowJo software.

In vitro IFNβ ELISA. Cells were seeded at a density of 10,000 cells per well in a 96-well plate. Mouse IFNβ (PBL Assay Science) was then added. After 6 h of incubation at 37 °C, the supernatant was aspirated from the wells to remove the mouse IFNβ. The wells were then gently washed once with medium. Fresh warm medium was then replaced in all the wells. After 72 h of incubation at 37 °C, the supernatant was collected, and the concentration of IFNβ was determined using the VeriKine Mouse Interferon Beta ELISA Kit (PBL Assay Science) or Mouse IFN-beta Quantikine ELISA Kit (R&D Systems).

Cytokine analysis from tumour lysate. Tumours were isolated from mice on day 12 after inoculation. One hundred milligrams of tissue was collected in a 2-ml round-bottom eppendorf tube with pre-chilled 500 μl of cell lysis buffer (ThermoFisher, EPX-99999-000) supplemented with 1 mM PMSF, cOmplete Protease Inhibitor Cocktail and PhosSTOP Phosphatase Inhibitor Cocktail tablets (Roche). Tissues were homogenized with 5 mm stainless-steel beads on a TissueLyser machine (Qiagen) at 25 Hz for 1 min and centrifuged at 16,000*g* for 10 min at 4 °C. Protein concentration was assessed by BCA assay (ThermoFisher Scientific), and tissues were normalized to 10 mg/ml. Lysate was then probed for IFNβ or IFNγ protein levels by ELISA (mouse IFN-beta Quantikine ELISA, mouse IFNγ Quantikine ELISA, R&D Systems).

Western blotting. Whole-cell lysates were prepared in either SDS lysis buffer (60 mM Tris HCl, 2% SDS, 10% glycerol, complete EDTA-free protease-inhibitor (Roche), and 500 U/ml benzonase nuclease (Novagen)) or RIPA Lysis and Extraction Buffer (ThermoFisher Scientific). For blots of ADAR1 and epistasis genes, cells were stimulated with 1,000 U/ml mouse IFNβ (PBL) overnight before collecting lysates unless otherwise noted. Samples were boiled at 100 °C and clarified by centrifugation. Protein concentration was measured with a BCA protein assay kit (Pierce). Between 30 and 150 μg protein was loaded onto 4–12% Bolt Bis-Tris Plus gels (Life Technologies) in MES buffer (Life Technologies). Protein was transferred to 0.45-mm nitrocellulose membranes (Bio-Rad). Membranes were blocked in Tris-buffered saline plus 0.1% Tween 20 (TBS-T) containing 5% non-fat dry milk for 1 h at room temperature followed by overnight incubation with primary antibody at 4 °C. For Extended Data Fig. 1a, membranes were washed with TBS-T and incubated with HRP-conjugated secondary antibodies for 1 h at room temperature. HRP was activated with Supersignal West Dura Extended Duration

© 2019 Springer Nature Limited. All rights reserved.

Substrate (Pierce) and visualized with a chemiluminscent detection system using Fuji ImageQuant LAS4000 (GE Healthcare Life Sciences). For all other figures, membranes were incubated with Odyssey Blocking Buffer (LI-COR), and IRDye 800CW or 680RD secondary antibodies. Membranes were then visualized using the Odyssey CLx scanner (LI-COR), then analysed using ImageJ, Image Studio Lite and Adobe Photoshop software.

Antibodies. Flow cytometry antibodies are listed in Supplementary Table 1 and immunohistochemistry antibodies are listed above. For western blotting, primary antibodies against ADAR1 (15.8.6, Santa Cruz Biotechnology), PKR (EPR19374, Abcam), RIG-I (D14G6, Cell Signaling Technology), MDA5 (D74E4, Cell Signaling Technology), STAT1 (p91, Polyclonal Goat IgG, R&D Systems), and MAVS (Rabbit Polyclonal IgG, ThermoFisher Scientific)were used. Peroxidase-conjugated secondary antibodies against rabbit IgG, mouse IgG or goat IgG were purchased from Jackson Laboratories. IRDye secondary antibodies against rabbit IgG, mouse IgG or goat IgG were purchased from LI-COR Biosciences.

Quantitative PCR. For each replicate, one million tumour cells were collected and resuspended in buffer RLT (Qiagen, 79216). RNA was extracted using an RNeasy Mini Kit (Qiagen, 74104) as per the manufacturer's instructions. RNA was converted to cDNA using the Improm-II Reverse Transcription System (Promega, A3800). qPCR reactions were carried out in 20 reaction volumes with 10 μl of TaqMan Gene Expression Master Mix (Thermo Fisher Scientific, 4369016), 5 μl nuclease-free H2O, 1 μl of each probe and 3 μl of each cDNA sample. The qPCR reaction was run using a ViiA 7 Real-Time PCR System (Thermo Fisher Scientific) in a 96-well plate. FAM-tagged targets were quantified using the $\Delta\Delta C_t$ method relative to β-actin (*Actb*), which was VIC-tagged.

CRISPR sgRNA sequences. *Adar* sgRNA 1 CACCGTCTGGATTCACAACTCCAGG
Adar sgRNA 2 CACCGTCACAGCCCTACCTTGCCA
Adar sgRNA 3 CACCGTGTGACTCTCAGAAATCAG
Adar sgRNA 4 ACCGTTCCAAGTCAATCAGCACTG
Adar sgRNA 5 CACCGCACACAGCAGGGGTACACCA
Adar sgRNA 6 CACCGTCCGTCAAGTACCAGATGGG
Ddx58 sgRNA 1 CACCGCGTTGGAGATGCTAAGACCG
Ddx58 sgRNA 2 CACCGTCCGCCAGAGATGAACGAAG
Eif2ak2 sgRNA 1 CACCGTGGCTACTCCGTGCATCTGG
Eif2ak2 sgRNA 2 CACCGCTCGTCTATGACAAGTAAT
Ifih1 sgRNA 1 CACCGCGTAGACGACATATTACCAG
Ifih1 sgRNA 2 CACCGACATAACAGCAACATGGGCA
Ifnar2 sgRNA 1 CACCGTACCAGAGGGTGTAGTTAG
Ifnar2 sgRNA 2 CACCACACAAGCTGAGGAGACCGA
Ifngr1 sgRNA 1 CACCCGACTTCAGGGTGAAATACG
Ifngr1 sgRNA 2 CACCGGTATTCCCAGCATACGACA
Mavs sgRNA 1 CACCGACTCCTCCAGACCAACTCCG
Mavs sgRNA 2 CACCGGTCACAACATCCCTGACCA
Stat1 sgRNA 1 CACCGATCATCTACAACTGTCTGA
Stat1 sgRNA 2 CACCGTACGATGACAGTTTCCCCA
Control sgRNA 1 CACCGCGAGGTATTCGGCTCCGCG
Control sgRNA 2 CACCGCTTTCACGGAGGTTCGACG
Control sgRNA 3 CACCATGTTGCAGTTCGGCTCGAT
Control sgRNA 4 CACCACGTGTAAGGCGAACGCCTT
Control sgRNA 5 CACCATTGTTCGACCGTCTACGGG

Statistics. Statistical tests employed with the number of replicates and independent experiments are listed in the text and figure legends. All graphs with error bars report mean ± s.e.m. values except where indicated. *t*-tests were two-tailed in all cases. For box-plot elements, the centre line represents the median value, box limits represent upper and lower quartiles and whiskers represent minimum and maximum values. PRISM was used for basic statistical analysis and plotting (http://www.graphpad.com), and the R language and programming environment (https://www.r-project.org) was used for the remainder of the statistical analysis. Multiple hypothesis testing correction was applied where multiple hypotheses were tested and is indicated by the use of FDR.

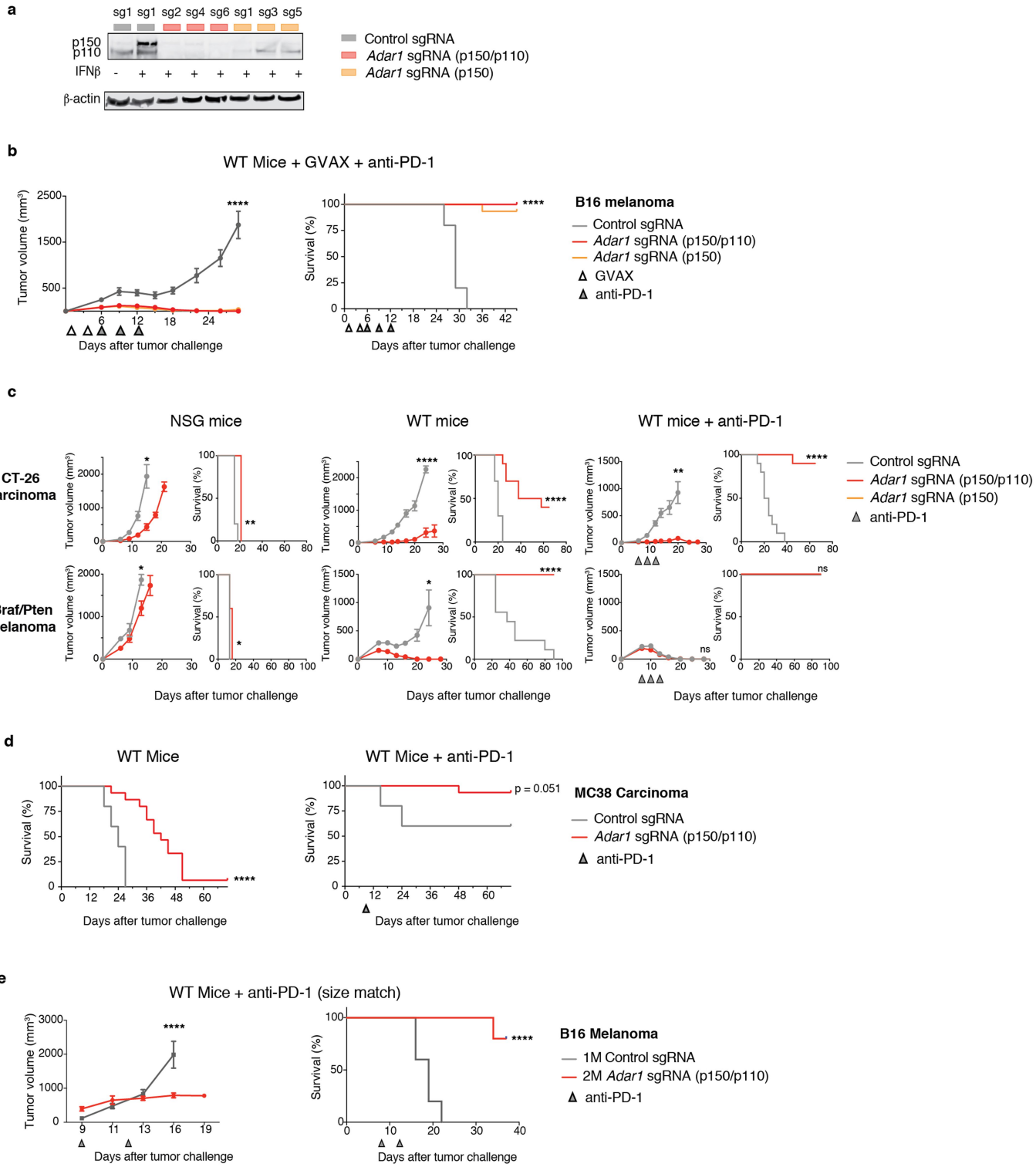
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data presented in this manuscript are available from the corresponding author upon reasonable request. Bulk tumour cell RNA sequencing has been deposited at the Gene Expression Omnibus (GEO) under accession number GSE110708. Single-cell RNA sequencing of tumour cells were also deposited at the GEO under accession number GSE110746.

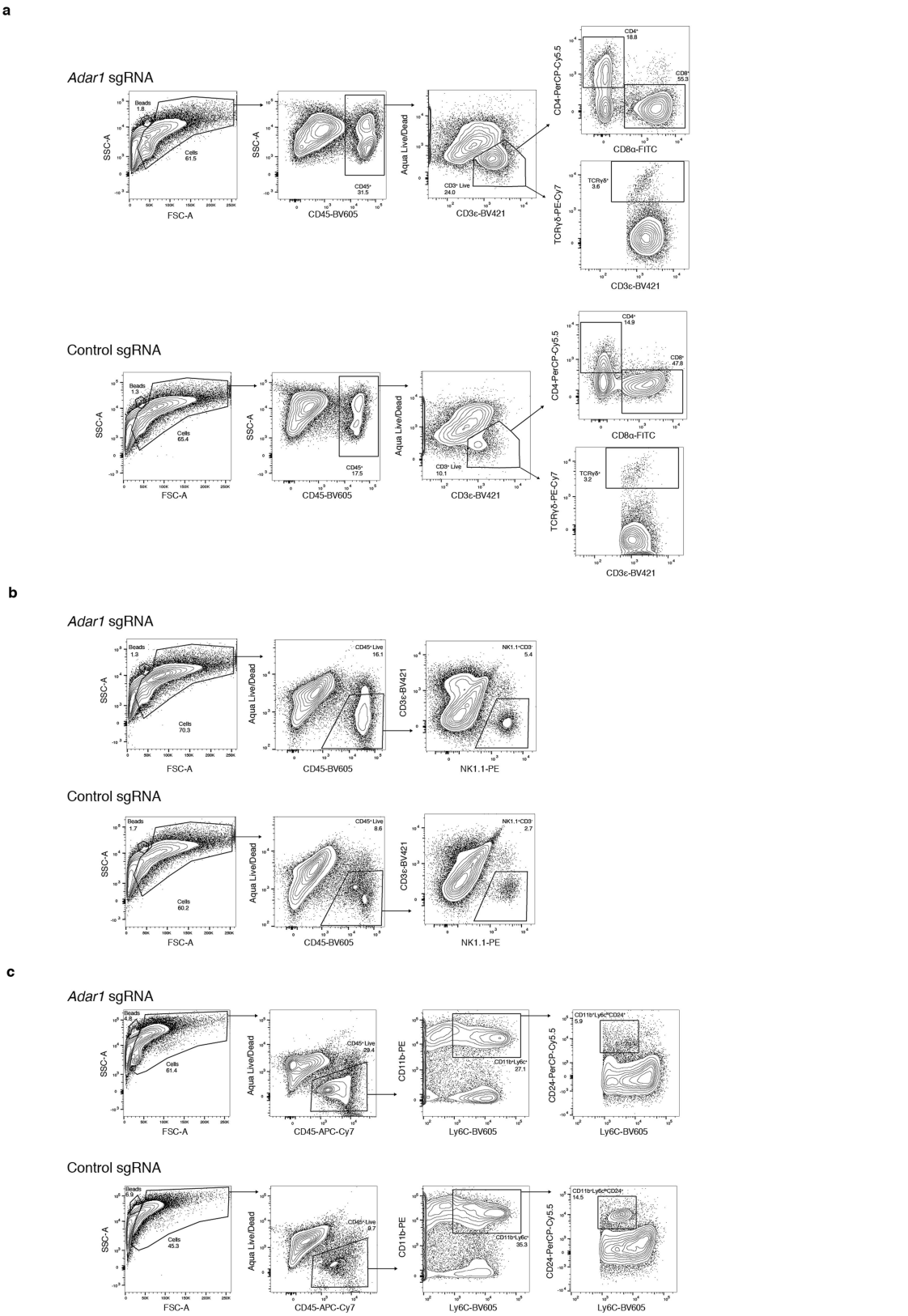
26. Ngwa, W., Tsiamas, P., Zygmanski, P., Makrigiorgos, G. M. & Berbeco, R. I. A multipurpose quality assurance phantom for the small animal radiation research platform (SARRP). *Phys. Med. Biol.* **57**, 2575–2586 (2012).
27. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat. Biotechnol.* **33**, 495–502 (2015).
28. Waltman, L. & van Eck, N. J. A smart local moving algorithm for large-scale modularity-based community detection. *Eur. Phys. J. B* **86**, 471 (2013).
29. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
30. DeTomaso, D. & Yosef, N. FastProject: a tool for low-dimensional analysis of single-cell RNA-Seq data. *BMC Bioinformatics* **17**, 315 (2016).
31. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.* **1**, 417–425 (2015).
32. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863–864 (2011).
33. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
34. Bazak, L. et al. A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. *Genome Res.* **24**, 365–376 (2014).
35. Porath, H. T., Carmi, S. & Levanon, E. Y. A genome-wide map of hyper-edited RNA reveals numerous new sites. *Nat. Commun.* **5**, 4726 (2014).
36. Yu, G., Wang, L.-G. & He, Q.-Y. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).
37. Paz-Yaacov, N. et al. Elevated RNA Editing activity is a major contributor to transcriptomic diversity in tumors. *Cell Reports* **13**, 267–276 (2015).
38. Newman, A. M. et al. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods* **12**, 453–457 (2015).
39. Yoshihara, K. et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat. Commun.* **4**, 2612 (2013).

© 2019 Springer Nature Limited. All rights reserved.



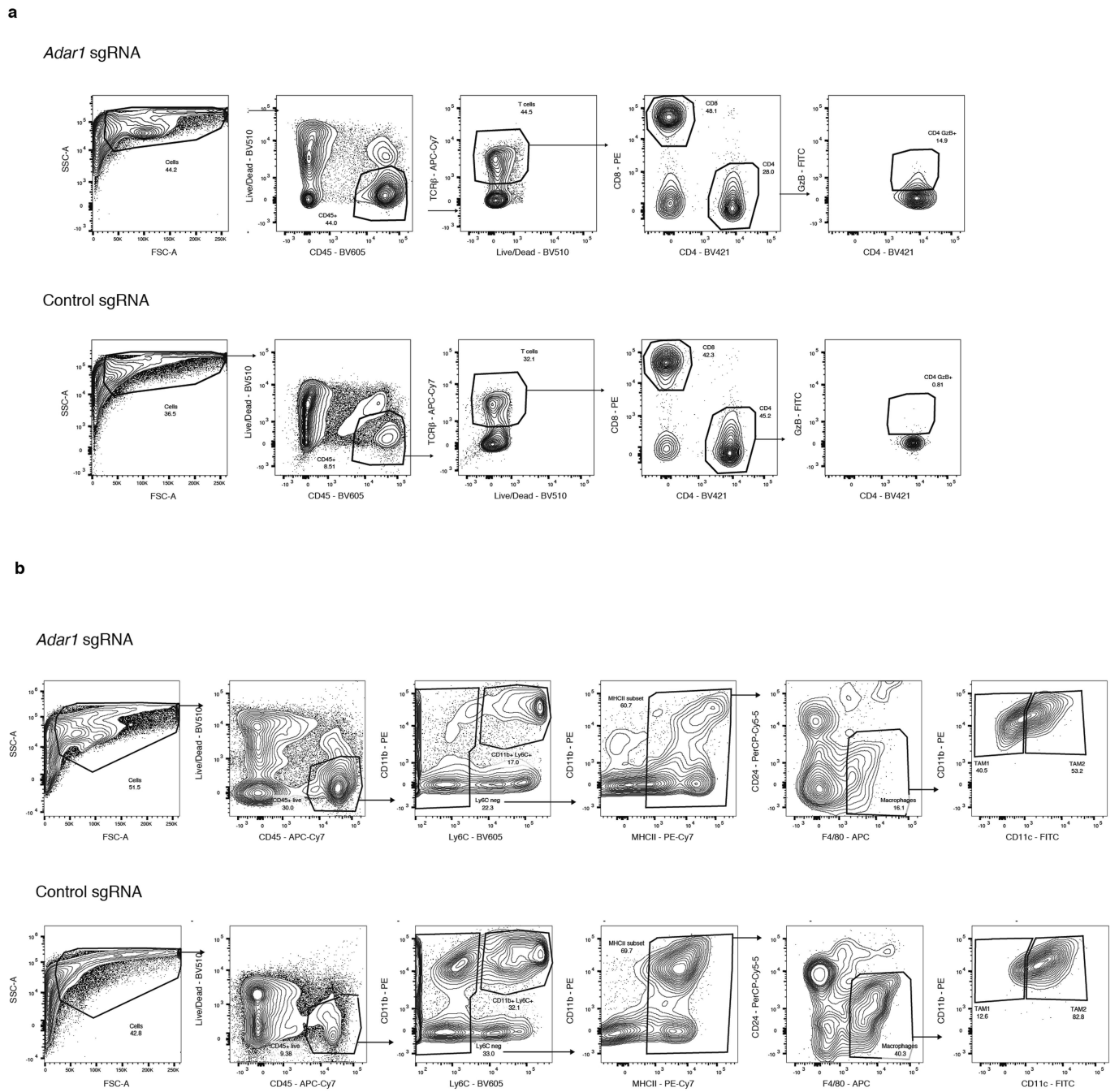
Extended Data Fig. 1 | Supporting evidence that ADAR1 loss enhances the response to immunotherapy. a, Expression of ADAR1 protein in control (grey), *Adar1* p150-null (orange) and *Adar1* p150/p110-null (red) B16 cells. Results are representative of three independent experiments. **b**, Tumour volume (left) and survival analysis (right) of control (grey), *Adar1* p150-null (orange) or *Adar1* p110/p150-null (red) B16 tumours in GVAX- and anti-PD-1-treated wild-type C57BL/6 mice. $n = 5$ animals per guide with two separate guides for the control group and at least two separate guides for each *Adar1*-null group. Data are representative of two independent experiments. **c**, Tumour volume and survival analysis of control (grey), *Adar1* p150-null (orange) or *Adar1* p110/p150-null (red) CT26 and *Braf/Pten* tumours in NSG, wild-type and wild-type anti-PD-1-treated mice. $n = 5$ mice per group; data are representative of two independent experiments. **d**, Survival analysis of control and *Adar1*-null MC38 tumours in wild-type and wild-type anti-PD-1-treated C57BL/6 mice. $n = 5$ animals per guide with two separate guides for the control group and three separate guides for the *Adar1*-null group. Data are representative of two independent experiments. **e**, Tumour volume and survival analysis of *Adar1*-null and control B16 tumours size matched at the time of PD-1 treatment initiation. **b**–**e**, Tumour volume curves are mean ± s.e.m and assessed with Student's *t*-test; survival curves assessed with log-rank test, $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$.

© 2019 Springer Nature Limited. All rights reserved.



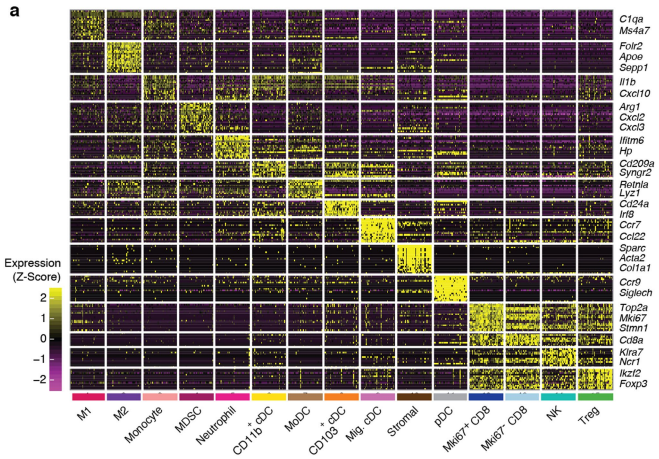
Extended Data Fig. 2 | Flow cytometry gating strategies and representative plots. a, Gating strategy and representative flow cytometry plots for the assessment of $CD4^+$, $CD8^+$ and γδ T cells in *Adar1*-null and control B16 tumours. **b**, Gating strategy and representative flow cytometry plots for the assessment of NK cells in *Adar1*-null and control B16 tumours. **c**, Gating strategy and representative flow cytometry plots for the assessment of $CD11b^+Ly6c^+$ and $CD11b^+Ly6c^{lo}CD24^+$ cells in *Adar1*-null and control B16 tumours.

© 2019 Springer Nature Limited. All rights reserved.



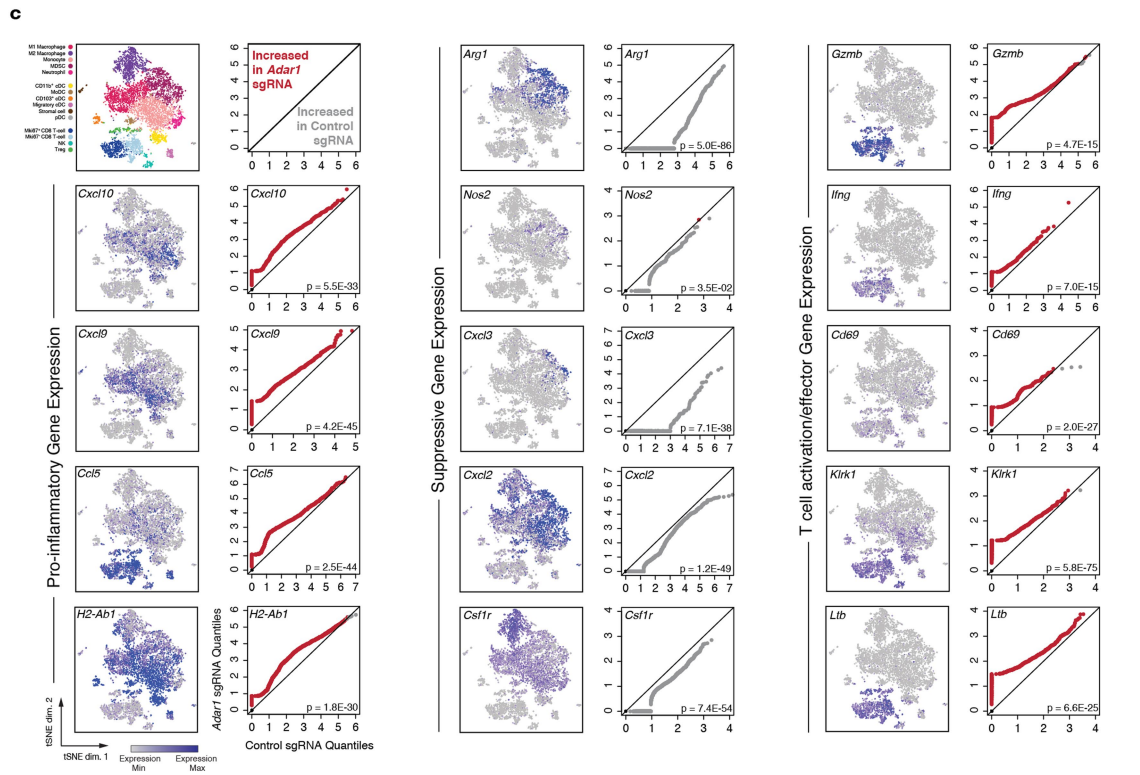
Extended Data Fig. 3 | Further flow cytometry gating strategies and representative plots. a, Gating strategy and representative flow cytometry plots for the assessment of granzyme $B^{+}CD4^{+}$ T cells in *Adar1*-null and control B16 tumours. **b**, Gating strategy and representative flow cytometry plots for the assessment of TAM1 and TAM2 populations in *Adar1*-null and control B16 tumours.

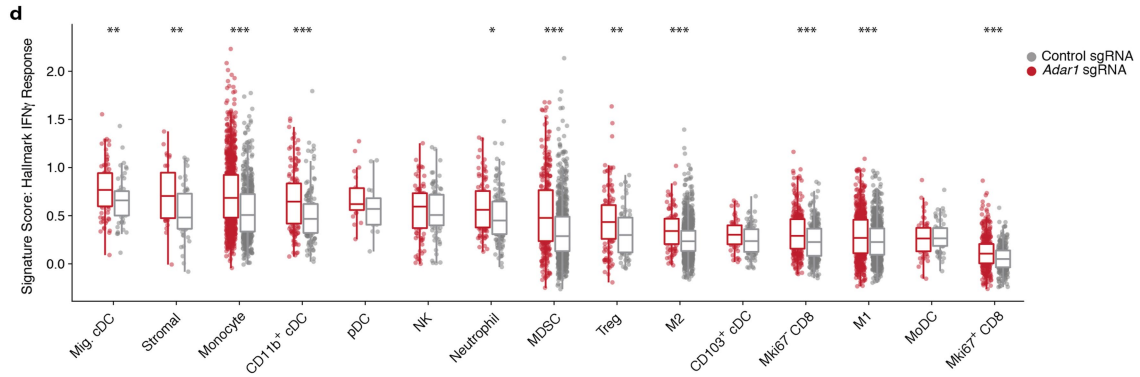
© 2019 Springer Nature Limited. All rights reserved.



b

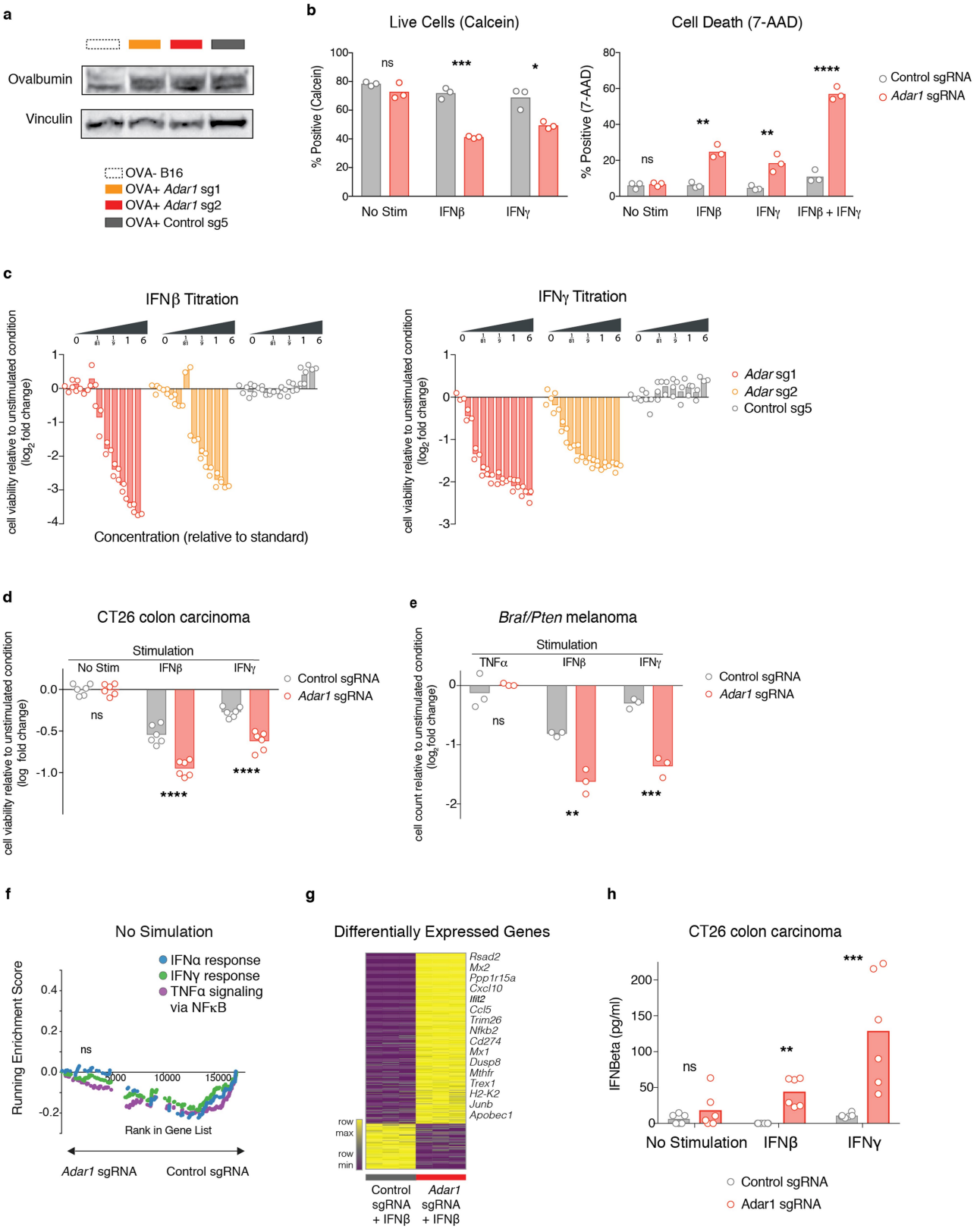
Population	Key Differentially Expressed Transcripts
M1 Macrophage	*Itgam, Adgre1, C1qa, Cxcl9, Vcam1*
M2 Macrophage	*Itgam, Adgre1, Mrc1, Clec10a, Folr2, Sepp1, Apoe, Maf*
Monocyte	*Itgam, Ly6c2, Il1b, Cxcl10, Ms4a4c*
MDSC	*Itgam, Arg1, Nos2, Pf4, Il1rn*
Neutrophil	*Itgam, Ly6c2, Hp, Msrb1, Ncf2*
CD11b+ cDC	*Itgax, Itgam, Zbtb46, Cd209a, H2-Ab1, H2-Eb1*
MoDC	*Itgam, Clec4b1, Ccl24, Ccr2, H2-DMa, H2-Ab1*
CD103+ cDC	*Flt3, Zbtb46, Batf3, Clec9a, Xcr1, Cd24a, Cd103, H2-Ab1*
Migratory cDC	*Flt3, Zbtb46, Batf3, Ccr7, Relb, Ccl22*
Stromal cell	*Sparc, Dcn, Igfbp1, Col3a1*
Plasmacytoid DC	*Ccr9, Siglech, Bst2*
Mki67+ CD8+ T cell	*Cd8a, Cd3g, Tox, Top2a, Cenpe, Cdk4, Thy1*
Mki67- CD8+ T cell	*CD8a, Cd3g, Thy1, Tox, Ccl5*
NK cell	*Ncr1, Klra7, Klre1*
Treg	*Cd4, Cd3g, Cd28, Ikzf2, Foxp3*





Extended Data Fig. 4 | Single-cell RNA-seq extended data. **a**, Gene expression matrix from single-cell RNA-seq experiment characterizing expression of lineage-defining genes in cell clusters. **b**, Key differentially expressed transcripts that distinguish cell clusters in Fig. 2. **c**, Paired quantile–quantile (Q–Q) plots comparing the expression of a curated set of genes in immune cells from *Adar1*-null and control tumours and matched *t*-SNEs depicting the distribution of gene expression for proinflammatory, suppressive and T cell activation/effector genes. *P* values calculated using Wilcoxon rank-sum test. **d**, Single-cell gene set enrichment scores of an IFNγ response signature score within individual immune subpopulations from *Adar1*-null and control tumours (*P* values calculated using Kolmogorov–Smirnov test). **a**, **c**, **d**, $n = 7{,}406$ cells. $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$.

© 2019 Springer Nature Limited. All rights reserved.

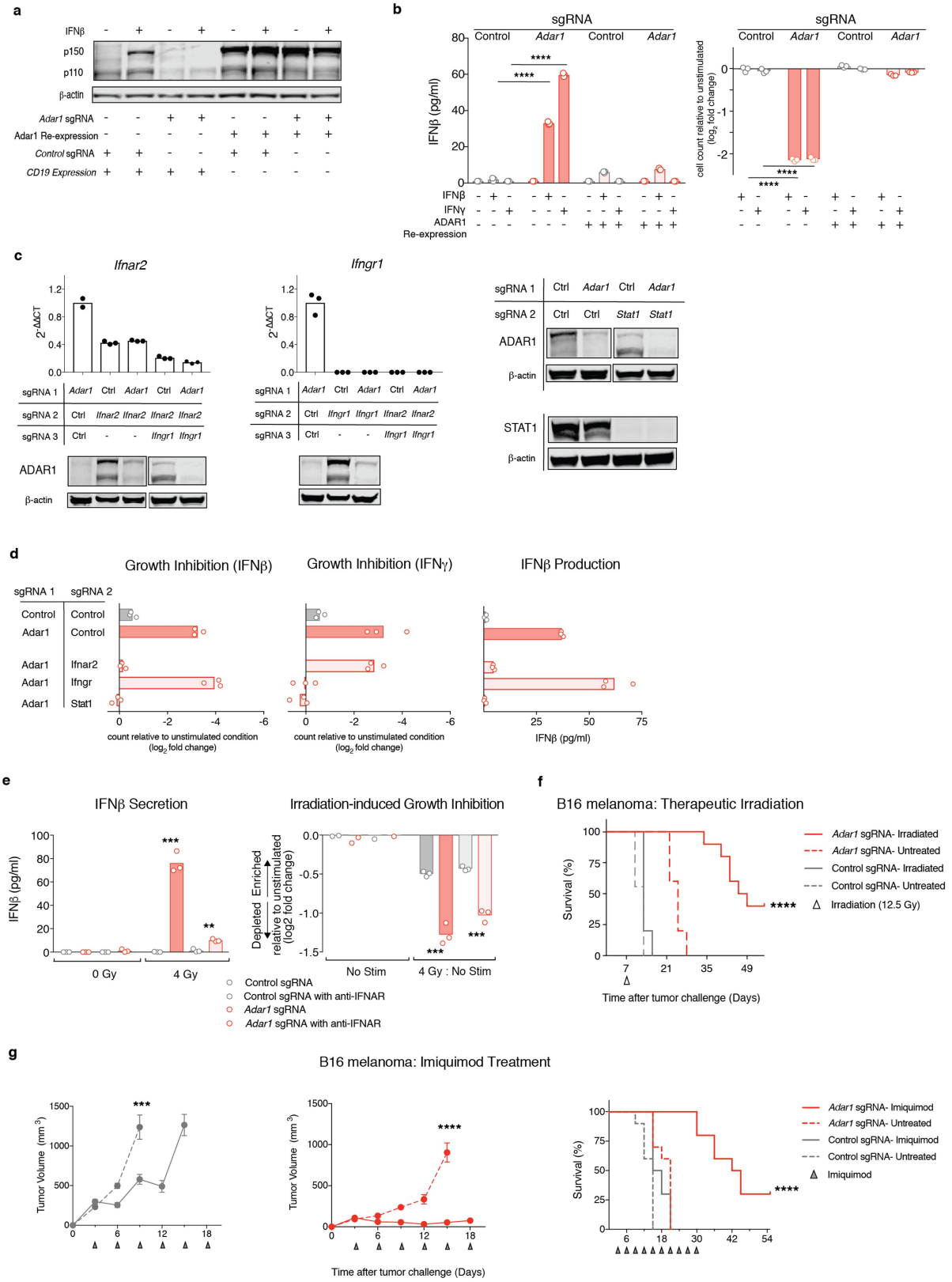


Extended Data Fig. 5 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Fig. 5 | Further studies corroborating the reported in vitro phenotype of *Adar1*-null tumour cells. a, Western blot demonstrating expression of ovalbumin in modified *Adar1* p150/p110-null (red), *Adar1* p150-null (orange) and control (grey) B16 tumour cell lines. Data are representative of two independent experiments. **b**, Calcein cell viability and 7-AAD cell death staining of control or *Adar1*-null B16 tumour cells following stimulation with IFNβ, IFNγ or a combination of both. Data are representative of three independent experiments with $n = 3$ for each condition. **c**, Growth and viability of *Adar1* p150/p110-null, *Adar1* p150-null and control B16 tumour cells in response to increasing doses of IFNβ and IFNγ ($n = 3$ for each condition). Doses are relative to 1× standard of 1,000 U ml^{-1} IFNβ and 100 ng ml^{-1} IFNγ. Data are representative of two independent experiments. **d**, Growth and viability of *Adar1* p150/p110-null and control CT26 tumour cells following stimulation with IFNβ or IFNγ relative to the unstimulated state ($n = 3$ for each condition). Data are representative of two independent experiments. **e**, Growth and viability of *Adar1* p150/p110-null and control *Braf*/*Pten* tumour cells following stimulation with TNF, IFNβ or IFNγ relative to the unstimulated state ($n = 3$ for each condition). Data are representative of two independent experiments. **f**, GSEA of gene signatures in *Adar1*-null compared with control B16 tumours cells after in vitro culture without cytokine stimulation. $n = 3$ for each condition; FDR calculated using GSEA. **g**, Heat map showing differentially expressed genes from *Adar1*-null and control B16 tumour cells 36 h after IFNβ stimulation in vitro ($n = 3$ for each condition). Genes listed in adjacent text were manually curated as antiviral or relevant to anti-tumour immunity. **h**, IFNβ ELISA of control and *Adar1* p150/p110 CT26 tumour cells following stimulation with IFNβ or IFNγ ($n = 3$ for each condition). **b**–**e**, **h**, Two-sided Student's *t*-test, $*P < 0.05$; $**P < 0.01$; $***P < 0.001$; $****P < 0.0001$.

© 2019 Springer Nature Limited. All rights reserved.

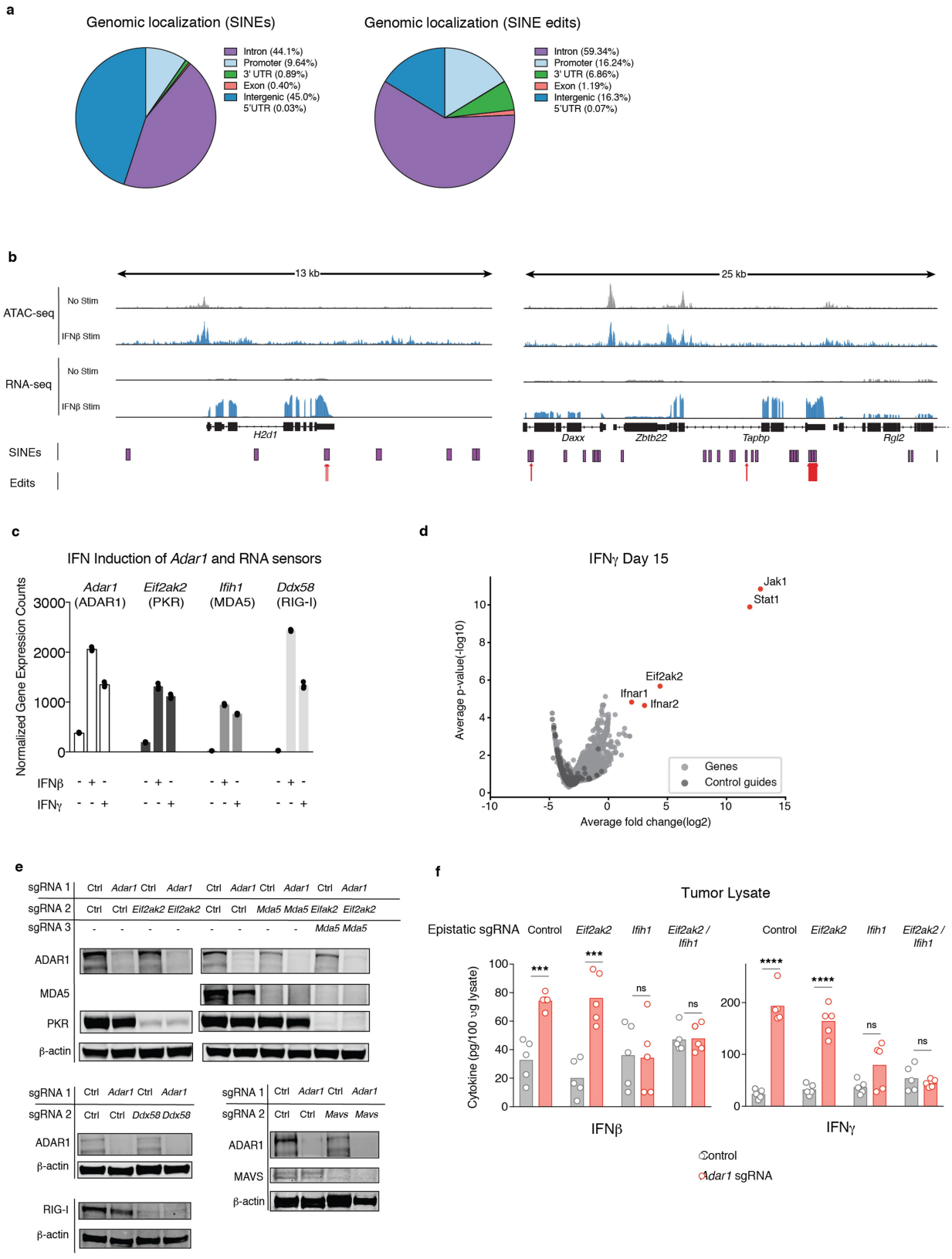


Extended Data Fig. 6 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Fig. 6 | ADAR1 re-expression and corroborating in vitro epistatic IFN-signalling experiments. a, Western blot of B16 *Adar1*-null tumour cells following re-expression of wild-type ADAR1 or an irrelevant control (CD19) protein. Data are representative of two independent experiments. **b**, IFNβ secretion (left) and relative growth (right) of control (grey), *Adar1* p150/p110-null (red), *Adar1*-null with full-length ADAR1 re-expression construct (red outline) and control with ADAR1 re-expression construct (grey outline) B16 tumour cells following cytokine stimulation as indicated ($n = 3$ for each condition). Data are representative of two independent experiments. **c**, qPCR and western blot validation of the loss of expression of *Ifnar2*, *Ifngr1* and *Stat1* from B16 tumour cells used to generate the control and *Adar1*-null tumour cell lines shown in Fig. 3. $n = 3$ for qPCR experiments and data are representative of two independent experiments. **d**, Growth inhibition (left two panels) and IFNβ ELISA (right panel) of control and *Adar1*-null B16 tumour cells modified to delete *Ifnar2*, *Ifngr1* or *Stat1* ($n = 3$ for each condition; data representative of three independent experiments). **e**, IFNβ secretion in vitro following irradiation with 4 Gy in *Adar1*-null and control B16 tumour cells with and without IFNAR-blocking antibodies (left). Growth and viability of *Adar1*-null and control B16 tumour cells in vitro following irradiation with 4 Gy with and without IFNAR-blocking antibodies (right). For both plots: $n = 3$ for each condition; data are representative of two independent experiments. **f**, Survival analysis corresponding to the tumour volume curves depicted in Fig. 3h of *Adar1* and control tumours treated with therapeutic irradiation. $n = 10$ mice for each group. Data are representative of two independent experiments. **g**, Tumour volume and survival analysis of control and *Adar1*-null B16 tumours treated with topical imiquimod. Data are representative of two independent experiments with $n = 10$ mice per group. **b**, **e** and tumour volume curves, two-sided Student's *t*-test; survival curves, log-rank test, $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$.

© 2019 Springer Nature Limited. All rights reserved.

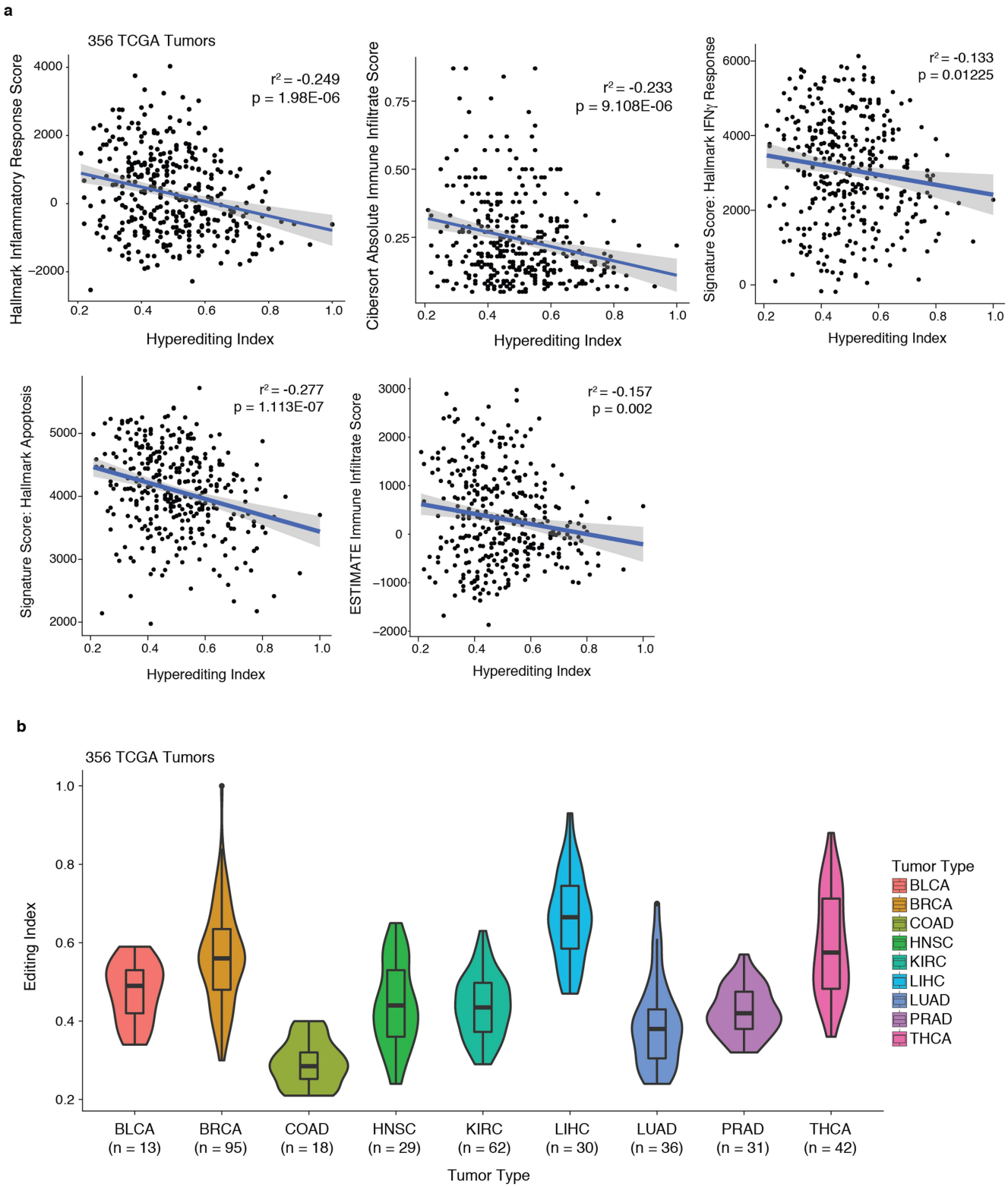


Extended Data Fig. 7 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

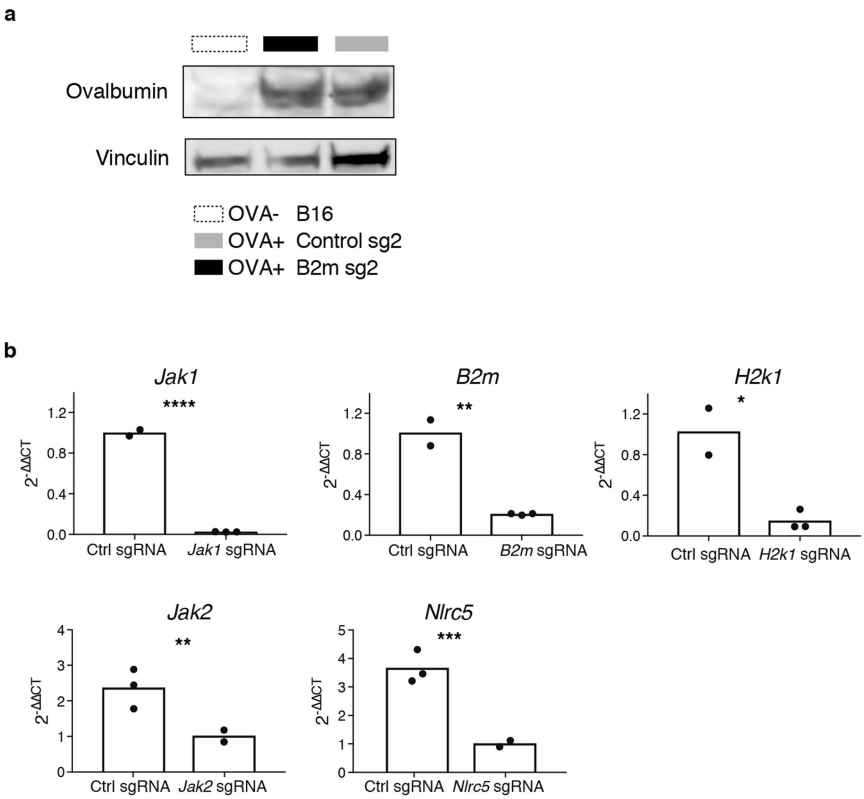
Extended Data Fig. 7 | Corroborating data for dsRNA editing and epistasis studies of *Adar1*-null tumours. a, Genomic localization of SINEs (left) and detected editing sites within SINEs (right) in control B16 tumour cells. **b**, Representative tracks of ATAC-seq and RNA-seq mapped to SINEs and detected edits in IFN-inducible regions of accessible chromatin and transcription. **c**, Transcriptional upregulation of *Adar1* and dsRNA sensors 36 h after stimulation with IFNβ or IFNγ in control B16 tumour cells as measured by RNA-seq ($n = 3$ for each condition). **d**, Volcano plot depicting the relative depletion and enrichment of sgRNAs targeting 20,146 genes in a $Cas9^+$ *Adar1*-null B16 tumour cell line following stimulation with IFNγ in vitro. *P* values are derived using STARS v1.3. **e**, Western blots demonstrating the loss of expression of PKR, MDA5, RIG-I, MAVS and ADAR1 from double knockout and triple knockout B16 tumour cell lines. Data are representative of two independent experiments. **f**, IFNβ and IFNγ ELISAs from tumour lysate extracted from *Adar1*-null and control tumours that were epistatically deleted for dsRNA sensors including *Eif2ak2* (PKR), *Ifih1* (MDA5) or both ($n = 5$ for each condition). **f**, Two-sided Student's *t*-test, $^*P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$.

© 2019 Springer Nature Limited. All rights reserved.



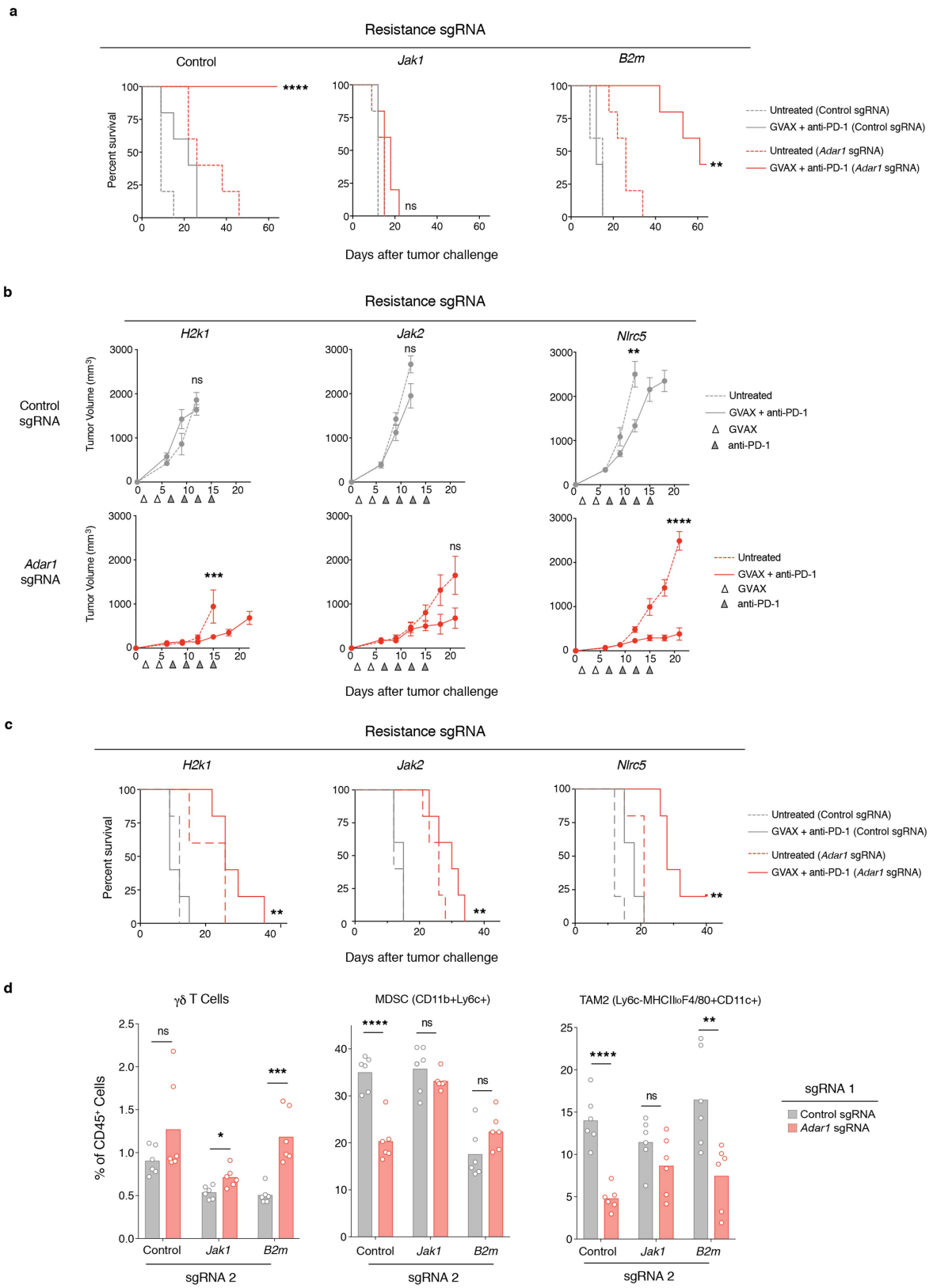
Extended Data Fig. 8 | Correlations between RNA editing and signatures of immune infiltration in TCGA. a, Pearson's correlations between hyperediting index and Hallmark Inflammatory Response, CIBERSORT Absolute Immune Infiltrate, Hallmark Interferon Gamma Response, Hallmark Apoptosis and ESTIMATE immune infiltrate gene signatures from 356 tumours in TCGA for which hyperediting index information was available. **b**, Distribution of hyperediting index values of individual tumour types from the same samples from TCGA. Box plots represent the range, median, 25th and 75th percentile with *n* as indicated in the figure.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Corroborating data for models of immunotherapy resistance. a, Western blot demonstrating the expression of ovalbumin in control and *B2m*-null B16 tumour cell lines depicted in Fig. 5c. Data are representative of two independent experiments. **b**, Quantitative PCR and western blots demonstrating loss of expression of *B2m*, *Jak1*, *H2k1*, *Jak2*, and *Nlrc* in B16 tumour cell lines used to make epistatically deleted *Adar1*-null or control tumour cells lines. $n = 3$ for qPCR experiments and data are representative of two independent experiments with P value calculated using two-sided Student's t-test, $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$; $^{****}P < 0.0001$.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Loss of ADAR1 overcomes multiple common mechanisms of resistance to immunotherapy in vivo. a, Survival analysis corresponding to tumour volume curves from Fig. 5c depicting the effect of *Adar1* deletion in B16 tumours in which *Jak1* or *B2m* have been epistatically deleted. $n = 5$ animals per group; data are representative of two independent experiments. **b**, Tumour volume curves in control (grey) and *Adar1-null* (red) B16 tumour cells epistatically deleted for *H2k1*, *Jak2* and *Nlrc5* and treated with GVAX and anti-PD-1 as indicated ($n = 5$ animals per group). **c**, Survival analysis corresponding to the tumour volume curves depicted in **b**. **d**, Additional TME characterization of control and *Adar1*-null B16 tumour cells epistatically deleted for *Jak1* and *B2m* as indicated ($n = 5$ mice per group). Data in **d** and all tumour volume curves assessed with two-sided Student's *t*-test; all survival curves assessed with log-rank test, $*P < 0.05$; $**P < 0.01$; $***P < 0.001$; $****P < 0.0001$.

© 2019 Springer Nature Limited. All rights reserved.

ARTICLE

https://doi.org/10.1038/s41586-018-0736-4

Cryo–EM structures and dynamics of substrate–engaged human 26S proteasome

Yuanchen Dong[1,2,3,4,10], Shuwen Zhang[1,5,10], Zhaolong Wu[1], Xuemei Li[6], Wei Li Wang[1,2,3,4], Yanan Zhu[1,5], Svetla Stoilova-McPhie[7], Ying Lu[8], Daniel Finley[9] & Youdong Mao[1,2,3,4,5,6]*

The proteasome is an ATP–dependent, 2.5–megadalton molecular machine that is responsible for selective protein degradation in eukaryotic cells. Here we present cryo–electron microscopy structures of the substrate–engaged human proteasome in seven conformational states at 2.8–3.6 Å resolution, captured during breakdown of a polyubiquitylated protein. These structures illuminate a spatiotemporal continuum of dynamic substrate–proteasome interactions from ubiquitin recognition to substrate translocation, during which ATP hydrolysis sequentially navigates through all six ATPases. There are three principal modes of coordinated hydrolysis, featuring hydrolytic events in two oppositely positioned ATPases, in two adjacent ATPases and in one ATPase at a time. These hydrolytic modes regulate deubiquitylation, initiation of translocation and processive unfolding of substrates, respectively. Hydrolysis of ATP powers a hinge–like motion in each ATPase that regulates its substrate interaction. Synchronization of ATP binding, ADP release and ATP hydrolysis in three adjacent ATPases drives rigid–body rotations of substrate–bound ATPases that are propagated unidirectionally in the ATPase ring and unfold the substrate.

The ubiquitin-proteasome pathway (UPP) has a central role in selective protein degradation in eukaryotic cells. It regulates myriad cellular processes, such as the cell cycle, apoptosis, the immune response, inflammation, and the response to proteotoxic stress[1–3]. Ubiquitylated substrates are recognized and degraded by the 2.5-megadalton 26S proteasome holoenzyme[3]. The holoenzyme is assembled from a barrel-shaped, proteolytically active 20S core particle (CP)[4] and two 19S regulatory particles (RPs)[5] capping both ends of the CP cylinder. The RP controls substrate access into the CP and is formed from the lid and base subcomplexes. Recognition of ubiquitylated substrates is mediated by the ubiquitin receptors RPN1, RPN10 and RPN13[2,3,6,7]. When a substrate is captured by the RP, the globular domains of the substrate are mechanically unfolded by a ring-like heterohexameric adenosine triphosphatase (ATPase) motor in the base. The motor module consists of six distinct subunits (RPT1–RPT6) from the 'ATPases associated with diverse cellular activities' (AAA) family[1–3], and regulates the engagement, RPN11-catalysed deubiquitylation[8] and degradation of substrates in an ATP-dependent manner[9] through hitherto unknown mechanisms.

Previous cryo-electron microscopy (cryo-EM) analyses have revealed the architecture of the substrate-free holoenzyme in six distinct states[3,10–22]. However, it remains unclear how the conformations of the substrate-free holoenzyme are related to its functional states in the presence of substrate. Here we describe atomic structures of the substrate-engaged human proteasome in seven conformational states. Our analysis reveals mechanisms by which the substrate is engaged, deubiquitylated, unfolded and translocated by the human proteasome.

Overview of seven conformational states

To capture the human proteasome in the action of substrate processing, we used the model substrate Sic1^{PY} (the Cdk inhibitor Sic1 from *Saccharomyces cerevisiae* with a PY motif, Pro-Pro-Pro-Ser, inserted into its N terminus)[7,23,24] and a nucleotide-substitution strategy. In brief, the purified holoenzyme was first primed with polyubiquitylated Sic1^{PY} in stoichiometric excess in the presence of 1 mM ATP for a short time, then supplied with 1 mM ATPγS (adenosine 5′-*O*-(3-thio)triphosphate) before being vitrified into cryo-EM samples (see Methods). The slowly hydrolysed ATPγS is expected to compete with ATP to occupy nucleotide-binding pockets in the AAA-ATPases, thus potentially pausing the substrate-engaged proteasome in any possible intermediate state before the completion of substrate degradation. Through this approach we determined cryo-EM structures of the substrate-engaged proteasome in seven distinct conformational states—designated E_{A1}, E_{A2}, E_B, E_{C1}, E_{C2}, E_{D1} and E_{D2}—to nominal resolutions of 2.8–3.6 Å (Fig. 1, Extended Data Figs. 1–3, Extended Data Table 1).

States E_{A1} and E_{A2} presumably represent conformations of initial ubiquitin recognition, and capture snapshots before and after, respectively, one ubiquitin near RPN10 is bound to RPN11[7,25] (Fig. 1f, Extended Data Figs. 3a, 4a, b). State E_B reveals a novel deubiquitylation-compatible complex, in which the isopeptide bond between RPN11-bound ubiquitin and substrate lysine is visible in the vicinity of the zinc-binding site of RPN11[11] (Fig. 1a, d). States E_{C1} and E_{C2} represent two successive conformations that are compatible with the initiation of substrate translocation, whereas states E_{D1} and E_{D2} capture two consecutive conformations of processive substrate translocation (Fig. 1b, c, Extended Data Fig. 3a).

Key structural features of these seven states show notable spatiotemporal continuity (Fig. 2, Extended Data Figs. 4–8, Extended Data Table 2, Supplementary Video 1). Foremost, in both states E_{A1} and E_{A2}, a ubiquitin density is observed on the T1 site of RPN1[6], and two ubiquitin densities—presumably linked via Lys63—abut the RPT4–RPT5 N-terminal coiled-coil (CC) domain next to RPN10[7,25,26] (Fig. 1f, Extended Data Fig. 4a). One ubiquitin is tightly bound to RPN11

[1]State Key Laboratory for Artificial Microstructures and Mesoscopic Physics, School of Physics, Peking University, Beijing, China. [2]Intel Parallel Computing Center for Structural Biology, Dana-Farber Cancer Institute, Boston, MA, USA. [3]Department of Cancer Immunology and Virology, Dana-Farber Cancer Institute, Boston, MA, USA. [4]Department of Microbiology and Immunobiology, Harvard Medical School, Boston, MA, USA. [5]Center for Quantitative Biology, Peking University, Beijing, China. [6]Electron Microscopy Laboratory, School of Physics, Peking University, Beijing, China. [7]Center for Nanoscale Systems, Harvard University, Cambridge, MA, USA. [8]Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [9]Department of Cell Biology, Harvard Medical School, Boston, MA, USA. [10]These authors contributed equally: Yuanchen Dong, Shuwen Zhang. *e-mail: youdong_mao@dfci.harvard.edu

© 2019 Springer Nature Limited. All rights reserved.

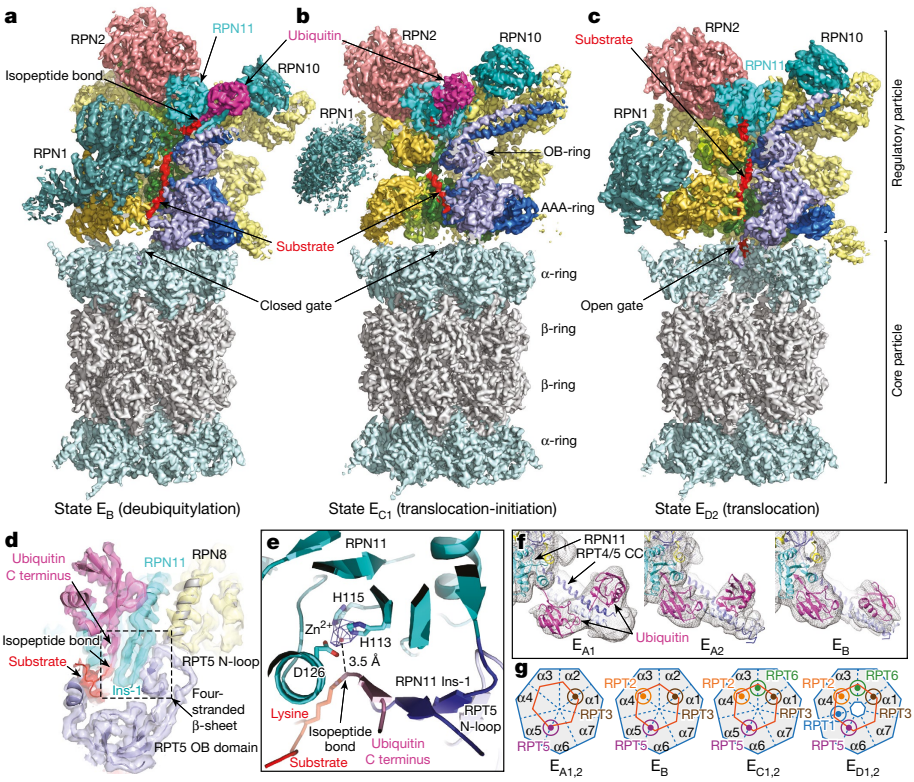


Fig. 1 | Cryo-EM structures of the substrate-bound human proteasome in distinct states. a–c, Cryo-EM density maps of substrate-bound human proteasome in state E_B at 3.3 Å (**a**), in state E_{C1} at 3.5 Å (**b**) and in state E_{D2} at 3.2 Å (**c**). The RPT1 density is omitted in **a–c** to show the substrate density inside the ATPase ring. Two α-subunits are omitted to show substrate density inside the CP gate in **c**. **d**, A close-up view of the quaternary interface around the scissile isopeptide bond between the RPN11-bound ubiquitin and the substrate lysine in E_B. The cryo-EM density is rendered as a transparent surface, superimposed with the cartoon representation of the atomic model. **e**, A close-up view of the zinc ion (hot pink sphere) closely approached by the isopeptide bond. The zinc ion density is shown as a blue mesh at 10σ level. Side chains of RPN11 that coordinate with the zinc ion are labelled. **f**, Comparison of two ubiquitin moieties between RPN11 and the CC domain of RPT4/5 among E_{A1}, E_{A2} and E_B. This observation is compatible with a recent single-molecule study[25]. The cryo-EM densities rendered as grey mesh representations are low-pass-filtered to 8 Å. The atomic model of ubiquitin is shown as a magenta cartoon representation. **g**, A schematic of RPT C-terminal insertion into the α-pockets of the CP and the state of the CP gate in all states (Extended Data Fig. 5). The CP is represented as a heptagon, the ATPase ring as a hexagon and the RPT C-tail insertion as a coloured sphere in a circle.

throughout states E_{A2}, E_B and E_{C1}, but it is released in E_{C2}, E_{D1} and E_{D2} (Fig. 1a–c, Extended Data Figs. 3a, 4b). Notably, two, three, four and five C-terminal tails of the ATPases are inserted into the inter-subunit surface pockets on the α-ring of the CP in E_A, E_B, E_C and E_D, respectively[10,11,27] (Fig. 1g, Extended Data Fig. 5). The CP gate is thus closed in $E_{A1,2}$, E_B and $E_{C1,2}$, but open in $E_{D1,2}$[10–13,27] (Extended Data Fig. 5a). States E_{A1} and E_{A2} also present 13 potential substrate densities in the chamber of the closed CP, including one in contact with the proteolytic site Thr1 of the β2-subunit[4], which is found in all states (Extended Data Fig. 6a, Extended Data Table 2). More details regarding the spatiotemporal continuity in substrate binding and nucleotide states are described in the following sections. Together, these observations indicate that the seven states are on the pathway of substrate processing by the holoenzyme.

Quaternary structure for substrate deubiquitylation

A prominent feature in state E_B is the formation of a quaternary subcomplex involving substrate-ubiquitin-bound RPN11, RPN8 and the N-loop (residues 99–119) of RPT5, which emanates from the top of its oligonucleotide- or oligosaccharide-binding (OB) domain[28] (Fig. 1d–f). To facilitate such a quaternary rearrangement, the lid is rotated outwards away from the axis of the OB ring relative to state E_A, which results in a widening of the entrance to the ATPase axial channel (Fig. 2a, Extended Data Fig. 4f, g).

The ubiquitin-bound RPN11–RPN8–RPT5 subcomplex starts to form in state E_{A2}, although the ATPase ring is not yet engaged with the substrate (Extended Data Fig. 4a, b). RPN11-bound ubiquitin also contacts the RPT5 CC domain (at residues 61–64) in E_{A2} and resides midway between its positions in states E_{A1} and E_B (Fig. 1f). In state E_{A1}, this ubiquitin abuts the RPT4–RPT5 CC domain in the vicinity of RPN11, but does not contact RPN11 (Fig. 1f). It is presumably covalently conjugated to substrate and linked to RPN10-bound ubiquitin moieties via Lys63[7,23,26]. The RPT4–RPT5 CC domain is shifted up during the E_{A1}-to-E_{A2} transition, shrinking the gap between RPN11 and the CC domain of RPT4–RPT5. These observations suggest that E_{A2} captures an intermediate state in which a ubiquitin is being transferred from the CC domain of RPT4–RPT5 to RPN11.

The RPN11–ubiquitin interface is centred on a hydrophobic pocket around Trp111 and Phe133 of RPN11, with the positioning of ubiquitin being comparable to that in the crystal structure of the ubiquitin-bound yeast Rpn11–Rpn8 complex[29]. The insert-1 (Ins1) region of RPN11 assumes a β-hairpin conformation, and pairs on one side with the C-terminal strand of ubiquitin and on the other side with a segment of the RPT5 N-loop, forming a four-stranded β-sheet (Fig. 1d). This β-sheet directs the ubiquitin C terminus towards the catalytic zinc-binding site in RPN11 and places the isopeptide bond within approximately 3.5 Å of the zinc ion, which has a strong density in our cryo-EM maps (Fig. 1e). Compared to the crystal structure of ubiquitin-bound Rpn11 from yeast[29] and the RPN11 structure in state E_{C1}, the Ins1 β-hairpin is rotated outwards by 5 Å in state E_B, allowing proper coordination of the isopeptide bond with the zinc ion (Extended Data Fig. 4d, e). This finding suggests that the conserved RPT5 N-loop stabilizes the RPN11–ubiquitin contact and optimizes the orientation of the

© 2019 Springer Nature Limited. All rights reserved.

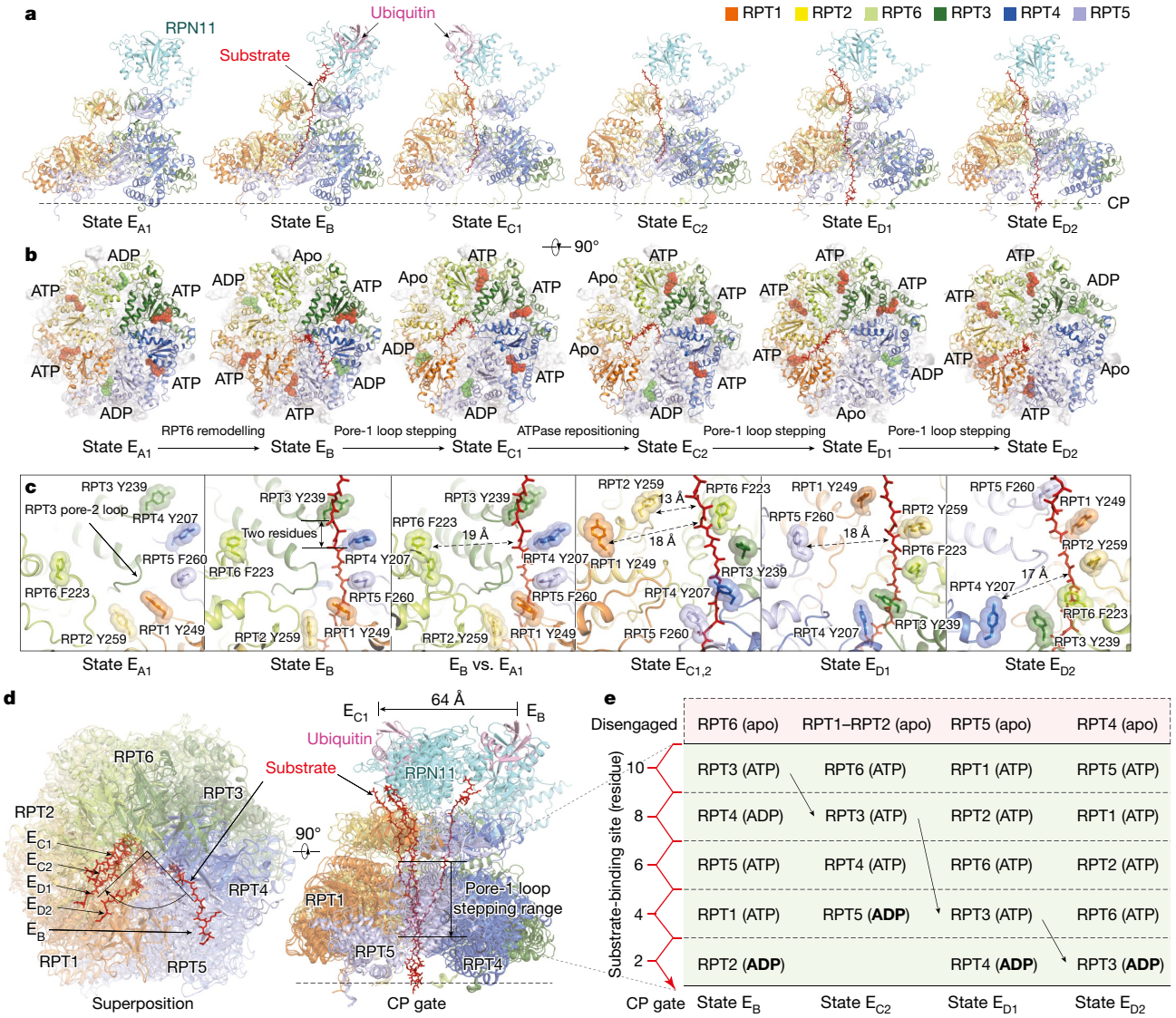


Fig. 2 | Dynamic substrate–proteasome interactions. a, Side views of the ATPase–RPN11 subcomplex interacting with the substrate in five states (E_B, $E_{C1,2}$ and $E_{D1,2}$) in comparison with state E_{A1}. The substrate is modelled as a polypeptide backbone structure and is represented with red sticks. Ubiquitin, RPN11 and the ATPases are rendered as transparent cartoons to show the substrate translocating inside the axial channel. The relative location of the CP is marked by the horizontal dashed line. Top right, colour codes of subunits used in all panels. **b**, Top views of the ATPase motors of distinct states. Nucleotides are shown in stick representation. The sphere representation of ADP and ATP is in green and in red, respectively. The CP is rendered as a grey surface representation. The structures are aligned together using their CP components. **c**, Varying architecture of pore-1 loop staircase interacting with substrate in five states (E_B, $E_{C1,2}$ and $E_{D1,2}$) as compared to that in state E_{A1}. Aromatic residues in the pore-1 loops are labelled and shown in stick representation superimposed with transparent sphere representation for highlighting. The distances from disengaged pore-1 loops to the substrate are marked. The third panel shows the superposition of state E_B with state E_A. **d**, Top view (left) and side view (right) of all substrate-bound ATPase–RPN11 structures superimposed together on the basis of structural alignment against the CP. **e**, A diagram summarizing the axial stepping of the substrate-contacting pore-1 loops that is coupled with nucleotide states. The vertical axis shows the relative locations of pore-1 loops interacting with the substrate, with the CP positioned at the bottom. Numbers label the relative distance from the lowest substrate–pore loop contact, using the number of residues as a metric. State E_{C1} is omitted here as its AAA-ATPase structure is identical to that of E_{C2}.

scissile isopeptide bond for efficient deubiquitylation. By contrast, the RPT5 N-loop is disordered in most other states (E_{A1}, $E_{C1,2}$ and $E_{D1,2}$) and the isopeptide bond is not observed between the RPN11-bound ubiquitin and substrate in E_{C1} (Fig. 1b).

The Ins1 region of RPN11 alternates among three distinct configurations (Extended Data Fig. 4c). It is a large open loop in state E_{A1}, folds into a β-hairpin throughout states E_{A2}, E_B and E_{C1} (whenever ubiquitin is bound), and is converted into a smaller, tighter loop in states E_{C2} and $E_{D1,2}$. The quaternary organization surrounding the zinc-binding site appears to explain why RPN11 exhibits much higher deubiquitylation activity in the context of the proteasome than in its non-proteasomal forms[5,30–32].

Substrate interactions with the holoenzyme

In the progression from state E_B to states $E_{D1,2}$, substrate contact with RPN11 is consistently centred around a hydrophobic groove at Phe118 and Trp121 of RPN11. In state E_B, this binding site faces the RPT3–RPT4 OB interface and the substrate extends straight from this site to Val125 of RPN11, beneath which the isopeptide bond linking the ubiquitin to the substrate lysine is held. In states $E_{C1,2}$ and $E_{D1,2}$, the substrate closely approaches Phe118 of RPT1 in the interior of the OB ring.

Within the axial channel of the ATPase ring, the substrate is threaded into a right-handed spiral staircase architecture in contact with the aromatic residues of pore-1 loops (Fig. 2a–c, Extended Data Fig. 6b–d). The aromatic side chains intercalate with the zigzagging main chain of

© 2019 Springer Nature Limited. All rights reserved.

the substrate through hydrophobic interactions. In addition, the main chains of the pore-1 loops potentially form hydrogen bonds with the main chain of the substrate. Substrate-contacting pore-1 loops are evenly distributed along the substrate, with two adjacent pore-1 loops spanning two amino acid residues in the substrate. The topology of this substrate–pore loop staircase architecture is nearly invariant from states E_B to E_{D2}. By contrast, the overall path of the translocating substrate changes markedly from E_B to $E_{C/D}$ (Fig. 2d).

The pore-1 loops of RPT3, RPT6, RPT1 and RPT5 reside at the highest position in contact with the substrate in states E_B, E_C, E_{D1} and E_{D2}, respectively. Notably, the pore-1 loop of RPT3 moves from the highest position to the lowest position in the substrate–pore loop staircase between E_B and E_{D2} (Fig. 2e). Meanwhile, the pore-2 loops support the opposite side of the substrate through their charged acidic residues, forming another, shorter staircase beneath that of the pore-1 loop. In states E_B, E_C, E_{D1} and E_{D2}, the pore loops from RPT6, the RPT1–RPT2 dimer, RPT5 and RPT4, respectively, are disengaged from the substrate (Fig. 2c).

A continuum of nucleotide states

The current resolution allows us to confidently distinguish ADP from ATP in the nucleotide-binding pockets of the ATPases (Fig. 2b, Extended Data Figs. 3c, 7). However, it is not possible to differentiate between ATP and ATPγS, as a mixture of both ATP and ATPγS was present in our buffer (see Methods). Thus, it remains possible that the nucleotide-binding sites that are competent for ATP hydrolysis are occupied by ATPγS in our structures. Previous studies have shown that ATPγS binding largely emulates ATP binding and may not necessarily change ATPase conformations[11]. For simplicity, we will refer to nonhydrolysed nucleotides as ATP hereafter.

Notably, the ADP-bound states navigate anticlockwise sequentially from RPT6 to RPT3 throughout all six ATPase subunits, indicating a full cycle of coordinated ATP hydrolysis throughout the AAA-ATPase ring from state E_A to state E_{D2} (Fig. 2b). The nucleotide states of the ATPases are strongly coupled with their substrate–pore loop interactions (Fig. 2e, Extended Data Figs. 7, 8a). Foremost, ATP is invariably found to bind the substrate-engaged ATPases at the middle or top positions in the pore-loop staircase, except in the early states E_A and E_B. Except for state E_A, at least one ATPase in each conformational state exhibits a very weak or partial density in its nucleotide-binding pocket (Extended Data Fig. 7). We refer to the nucleotide states of these ATPases as an apo-like state. The apo-like state is always observed in the ATPases the pore loops of which are disengaged from the substrate. Thus, all apo-like subunits form prominent gaps at the interfaces with their nearest neighbours on both sides, resembling in this respect RPT6 in state E_B (Fig. 3a, Extended Data Fig. 8a).

Adjacent to the apo-like subunit, the ATPase the pore-1 loop of which resides at the lowest position in contact with the substrate is always bound to ADP (Fig. 2b, e). The apo-like subunits and their anticlockwise adjacent ADP-bound subunits consistently exhibit a relatively wide nucleotide-binding pocket, with the arginine fingers from the clockwise adjacent subunit displaced more than 10 Å away from the Walker A motif (Extended Data Fig. 8a). By contrast, whenever the ATPases engaged with the substrate have their pore-1 loops located in the middle or top register in the substrate–pore loop staircase, the pocket is always tightly packed around magnesium-ion-bound ATP, with the *trans*-acting arginine fingers residing within 3–4 Å of either γ-phosphate or β-phosphate, indicating potential competence for ATP hydrolysis in these sites[33] (Extended Data Fig. 8a).

ATP-dependent substrate engagement

The organization of the substrate-engaged pore-1 loop staircase in state E_B is highly similar to that of the substrate-free pore-1 loop staircase in state E_A, suggesting that state E_B reflects substrate engagement before translocation (Fig. 2c). Structural comparison of states E_A and E_B clarifies how ATP hydrolysis regulates substrate engagement for deubiquitylation (Fig. 3). In state E_A, the AAA channel is too narrow to engage

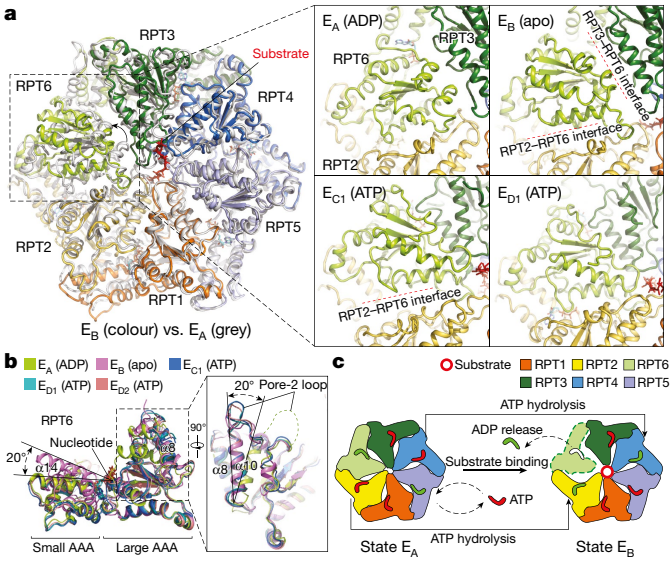


Fig. 3 | Structural basis for nucleotide-driven substrate engagement in the AAA-ATPase channel. a, Superposition of the AAA-ring structures of states E_A (grey) and E_B (colour). The insets show side-by-side comparisons of RPT6 conformations in the four most distant states. Interfacial gaps are marked with red dashed lines. **b**, Superposition of the RPT6 AAA domain structures from five distinct states aligned against the large AAA subdomain shows that RPT6 assumes three major conformations. Transition from E_A to E_B involves both refolding of the pore-2 loop (right insert) and a 20° rigid-body rotation between the large and small AAA subdomains. **c**, Schematic of ATP hydrolysis and nucleotide exchange required to implement the transition from E_A to E_B.

substrate, suggesting that it is in a closed state. To open the channel for substrate engagement, the AAA-ATPase ring must rearrange its quaternary organization.

Indeed, the AAA domain of RPT6 undergoes marked structural changes from E_A to E_B. An approximately 40° rotation out of the plane of the ATPase ring is observed in the large AAA subdomain of RPT6, creating prominent gaps at the RPT2–RPT6 and RPT3–RPT6 interfaces (Fig. 3a). By contrast, the AAA domains of other ATPases mostly move as a single rigid body with subtle changes restricted to the pore loops. In state E_A, the pore-2 loop of RPT6 is largely disordered. Notably, this loop refolds into an ordered structure in E_B, spanning residues 251–266 (Fig. 3b). Consistent with this observation, the ADP bound to RPT6 in state E_A is released in E_B. Thus, ATP hydrolysis and ADP release in RPT6 are programmed to trigger an iris-like movement in the ATPase ring that opens the axial channel (Fig. 2b, c). The coordinated ATP hydrolysis in RPT5 directly opposite from RPT6 in E_A, and in RPT4 opposite from RPT2 in E_B, is expected to increase the conformational flexibility of the ATPase ring that is required to open the AAA channel, thus allowing substrate engagement (Fig. 3c). Indeed, a greater degree of pore-1 loop movement is observed in RPT3, RPT4 and RPT5 than in RPT1 and RPT2 during the E_A-to-E_B transition (Fig. 3a).

Initiation of substrate translocation

From state E_A to state E_{D1}, the RPT1–RPT2 dimer undergoes a complete cycle of ATP hydrolysis and nucleotide exchange that initiates substrate translocation (Fig. 4). During the E_B-to-E_{C1} transition, an approximately 40° vertical rotation of the RPT1–RPT2 dimer moves the corresponding pore loops from the bottom of the substrate-pore loop staircase to almost the highest altitude, but 13–18 Å away from the substrate (Figs. 2c, 4a, Extended Data Fig. 4i). Concomitantly, binding of ATP to RPT6 promotes engagement of the RPT6 pore-1 loop with the substrate at the top of the substrate–pore loop staircase. Together, these conformational changes result in a one-step forward translocation of the substrate by a distance of two residues (Figs. 2e, 4b). Release of ADP in RPT1 during the E_{C1}-to-E_{C2} transition does not trigger

© 2019 Springer Nature Limited. All rights reserved.

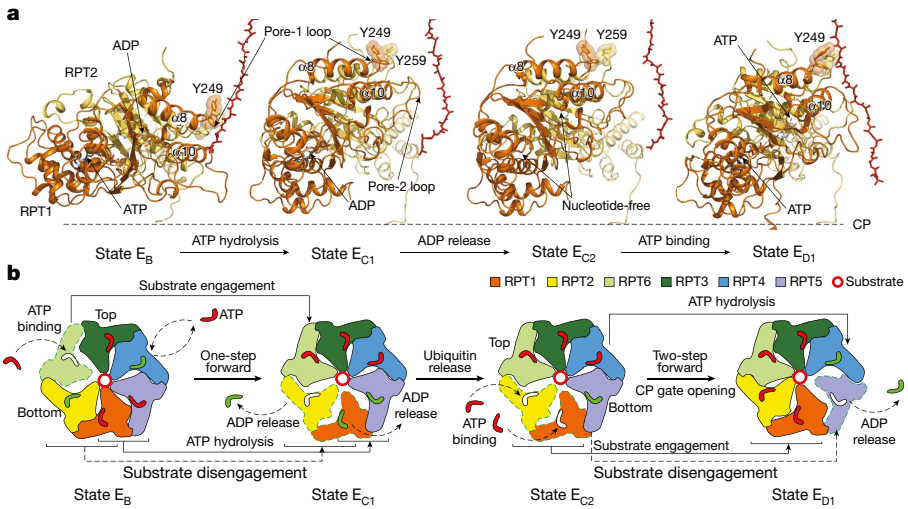


Fig. 4 | ATP hydrolysis drives initial three steps of substrate translocation through the AAA-ATPase channel. a, Side-by-side comparison of the RPT1–RPT2 dimer conformations in four sequential states that undergo a complete cycle of ATP hydrolysis and exchange in the RPT1–RPT2 dimer. Structures are aligned against the CP to show their conformational changes relative to the CP gate. Tyrosine residues in the pore-1 loops are shown in stick representation and highlighted with transparent sphere representation. The substrate is shown in red in stick representation. **b**, Schematic of the cascade of ATP hydrolysis and nucleotide exchange required to implement the transition from E_B to E_{D1}.

obvious conformational changes (Fig. 4a). However, during the E_{C2}-to-E_{D1} transition, binding of ATP to both RPT1 and RPT2 drives their substrate re-engagement at the top of the substrate–pore loop staircase and, together with ADP release from RPT5, results in a two-step forward translocation of the substrate (Fig. 4b). The total distance of substrate translation over the first three steps is comparable to that from the lowest substrate-contacting pore loop to the CP gate.

Notably, the conformation of the ATPase ring is nearly identical in states E_{C1} and E_{C2}, although multiple events are associated with the E_{C1}-to-E_{C2} transition, including ubiquitin dissociation from RPN11, ADP release from RPT1, conformational changes in the lid, and repositioning of the ATPase ring above the CP (Fig. 2a, b, Extended Data Fig. 4g, h). Thus, the initiation of substrate translocation is extensively coordinated with other regulatory events that prepare the proteasome for processive substrate degradation (Fig. 1g, Extended Data Table 2).

Substrate unfolding and translocation

Systematic structural alignment defines a generic hinge-like rotation of 15–25° in each ATPase between its small and large AAA subdomains upon nucleotide binding or release (Figs. 3b, 5b, Extended Data Fig. 8b–e). By contrast, the dihedral angle between the small and large AAA subdomains remains nearly invariant between the ADP-bound and ATP-bound states during substrate translocation, suggesting that nucleotide binding locks the AAA domain into a single rigid body. Hence, the release of γ-phosphate after ATP hydrolysis appears to be insufficient to immediately trigger intrinsic motion in the AAA domain; instead, it potentiates such conformational changes, which can be triggered later by either nucleotide exchange or changes in inter-ATPase interactions (Fig. 5a–c, Supplementary Discussion).

Structural comparison of states E_{D1} and E_{D2} provides insight into the mechanism of substrate unfolding and translocation (Fig. 5d). To establish processivity in substrate translocation, at least three adjacent ATPases must synchronize their nucleotide processing: the first binding an ATP, the second releasing an ADP and the third hydrolysing an ATP (Fig. 5d, left). In the second ATPase, release of ADP allows potential energy harvested from ATP hydrolysis to be converted into kinetic energy, which powers the most prominent vertical rotation (30–40°) in the ATPase ring and is transferred to adjacent ATPases from both sides along the ATPase ring (Fig. 5d, right; Extended Data Fig. 4i–l). On the clockwise side, the vertical rotation of the second ATPase away from the bottom of the substrate-pore loop staircase is synchronized with conformational changes in the first ATPase during its binding of ATP, which is coupled with substrate re-engagement of the first ATPase at the top of the substrate-pore loop staircase (Fig. 5c, d). On the anticlockwise side, the rotation facilitates ATP hydrolysis in the third ATPase, probably by repositioning the arginine fingers of the second ATPase that coordinate the neighbouring ATP[33]. Following ATP hydrolysis, release of γ-phosphate from the third ATPase abolishes the interaction between γ-phosphate and the *trans*-acting arginine fingers from the second ATPase, thus destabilizing their inter-subunit association and promoting disengagement of the second ATPase from the substrate at the bottom of the substrate-pore loop staircase (Fig. 5d). Such coordinated hydrolysis is essential for unidirectional propagation of conformational changes in the ATPase ring (Fig. 5c).

The large vertical rotation of the second ATPase away from the lowest register of the substrate–pore loop staircase makes room at the bottom and—via inter-subunit interactions—drives each of the four substrate-engaged ATPases to rotate downwards as a single rigid body with a small differential angle (5–10°; Fig. 5d, right). Together, these substrate-engaged ATPases collectively translate the substrate towards the CP via the pore-loop staircase. Through this highly concerted process, the chemical energy of ATP hydrolysis is converted into the mechanical work of substrate unfolding (Supplementary Video 2).

Long-range quaternary allosteric regulation

In both states E_{D1} and E_{D2}, the toroidal domain of RPN1 forms a surface cavity with the CC domain of RPT1–RPT2, into which a short helix of RPN2 is inserted (Extended Data Fig. 9a–c). This helix resides in the middle of a long loop (residues 820–871) that emanates from the RPN2 toroidal domain[34]. This long-range association of RPN1–RPN2—which is not observed in other states (E_A to E_C)—seems to stabilize a larger interface formed between RPN1–RPN2 and RPT1–RPT2, and may thus regulate substrate translocation allosterically. Furthermore, the CC domain of RPT5 switches its interaction between RPN9 and RPN10 in states E_{D1} and E_{D2}, which also seems to regulate substrate processing in a long-range fashion, consistent with a recent biochemical study[35] (Extended Data Fig. 9d, e, Supplementary Discussion).

Insights into the complete substrate-processing cycle

Our data suggest that ATP hydrolysis in the proteasome holoenzyme follows distinct modes at different stages of substrate processing (Fig. 6, Extended Data Fig. 10). The first mode features coordinated ATP hydrolysis in a pair of oppositely positioned ATPases. We speculate that this is orchestrated to promote initial substrate recognition and

© 2019 Springer Nature Limited. All rights reserved.

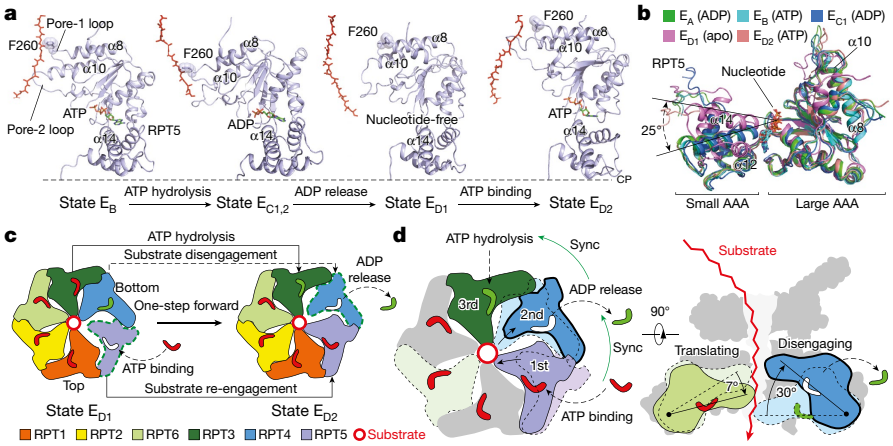


Fig. 5 | Mechanism for processive substrate translocation driven by a complete cycle of ATP hydrolysis. a, Side-by-side comparisons of RPT5 conformations in four sequential states that undergo a complete cycle of ATP hydrolysis and exchange in RPT5. The structures are aligned against the CP to show their conformational changes relative to the CP gate. Phenylalanine 260 of the pore-1 loop is shown in stick representation and highlighted with transparent sphere representation. The substrate is shown in red in stick representation. ADP release is coupled to disengagement of RPT5 pore-1 loop from the substrate at the bottom of the substrate-pore loop staircase during the E_{C2}-to-E_{D1} transition. ATP binding is coupled to re-engagement of RPT5 pore-1 loop with the substrate at the top of the substrate-pore loop staircase during the E_{D1}-to-E_{D2} transition. **b**, Superposition of RPT5 structures from five distinct states aligned against the large AAA subdomain shows that RPT5 assumes two major conformations between apo-like and nucleotide-bound states. **c**, Schematic of ATP hydrolysis and nucleotide exchange required to implement the transition from E_{D1} to E_{D2}. **d**, Schematic illustrates mechanism for processive substrate translocation. Synchronization of nucleotide processing in three adjacent ATPases (ATP binding, ADP release and ATP hydrolysis; left) creates differential vertical rigid-body rotations in each substrate-engaged ATPase that cooperatively translate the substrate (right).

deubiquitylation. This mode is reminiscent of the nucleotide-binding pattern observed in state S_{D2} of the substrate-free human proteasome[11] and the hexameric ClpX protease of *Escherichia coli*[36]. The second mode features coordinated ATP hydrolysis in two adjacent ATPases and is used to initiate substrate translocation and to coordinate CP gating. This mode seems to be compatible with previous studies that have suggested that binding of four nucleotides promotes substrate engagement[9,37,38]. Substrate processing culminates in a third hydrolytic mode, which is by comparison greatly simplified, as it features sequential hydrolysis of one nucleotide at a time. The third mode is likely to be the most efficient for maintaining processivity in substrate unfolding and translocation. In both the second and third modes, synchronized ATP binding and ADP release in at least two adjacent ATPases convert the chemical energy of ATP hydrolysis into differential rigid-body rotations in the ATPase ring that mechanically unfold the substrate and are propagated unidirectionally through coordinated ATP hydrolysis in the anticlockwise adjacent ATPase. The third mode is consistent with the proposed ATP hydrolysis mechanism of several other hexameric ATPase motors[39–46].

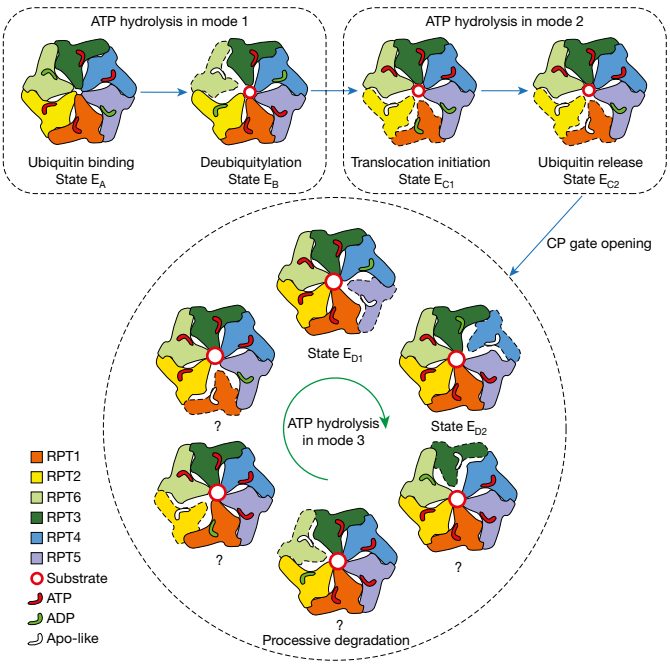


Fig. 6 | Model of the complete cycle of substrate processing by the human 26S proteasome. The cartoon summarizes the concept of three principal modes of coordinated ATP hydrolysis observed in the seven states and our proposal of how they regulate the complete cycle of substrate processing by the proteasome holoenzyme.

During transitions between consecutive states of the proteasome, the multiplicity of nucleotide-processing events in distinct ATPases implies that fast steps and sparsely populated intermediate states might have been missed in our cryo-EM reconstructions. Future studies to identify these missing intermediates will be required to clarify how ATP hydrolytic events and nucleotide exchange are coordinated with each other and allosterically linked to substrate translocation.

In summary, we have determined the atomic structures of the substrate-engaged human proteasome in seven native states during degradation of a polyubiquitylated substrate. These structures establish a foundation for understanding dynamic substrate–proteasome interactions during the complete cycle of substrate processing, and provide a wealth of atomic-level information accounting for several decades of biochemical studies of proteasome function[1–3,6–13,23–25,29,35,38]. A plethora of potential substrate-binding sites revealed in this study may facilitate the future development of drugs that modulate proteasome functions, which have been implicated in various diseases, such as multiple myeloma and neurodegenerative diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0736-4.

Received: 12 September 2018; Accepted: 2 November 2018;
Published online 12 November 2018.

1. Livneh, I., Cohen-Kaplan, V., Cohen-Rosenzweig, C., Avni, N. & Ciechanover, A. The life cycle of the 26S proteasome: from birth, through regulation and function, and onto its death. *Cell Res.* **26**, 869–885 (2016).

© 2019 Springer Nature Limited. All rights reserved.

2. Finley, D., Chen, X. & Walters, K. J. Gates, channels, and switches: elements of the proteasome machine. *Trends Biochem. Sci.* **41**, 77–93 (2016).
3. Bard, J. A. M. et al. Structure and function of the 26S proteasome. *Annu. Rev. Biochem.* **87**, 697–724 (2018).
4. Groll, M. et al. Structure of 20S proteasome from yeast at 2.4 Å resolution. *Nature* **386**, 463-471 (1997).
5. Lu, Y. et al. Conformational landscape of the p28-bound human proteasome regulatory particle. *Mol. Cell* **67**, 322–333.e6 (2017).
6. Shi, Y. et al. Rpn1 provides adjacent receptor sites for substrate binding and deubiquitination by the proteasome. *Science* **351**, aad9421 (2016).
7. Verma, R., Oania, R., Graumann, J. & Deshaies, R. J. Multiubiquitin chain receptors define a layer of substrate selectivity in the ubiquitin-proteasome system. *Cell* **118**, 99–110 (2004).
8. Verma, R. et al. Role of Rpn11 metalloprotease in deubiquitination and degradation by the 26S proteasome. *Science* **298**, 611–615 (2002).
9. Smith, D. M., Fraga, H., Reis, C., Kafri, G. & Goldberg, A. L. ATP binds to proteasomal ATPases in pairs with distinct functional effects, implying an ordered reaction cycle. *Cell* **144**, 526–538 (2011).
10. Chen, S. et al. Structural basis for dynamic regulation of the human 26S proteasome. *Proc. Natl Acad. Sci. USA* **113**, 12991–12996 (2016).
11. Zhu, Y. et al. Structural mechanism for nucleotide-driven remodeling of the AAA-ATPase unfoldase in the activated human 26S proteasome. *Nat. Commun.* **9**, 1360 (2018).
12. Eisele, M. R. et al. Expanded coverage of the 26S proteasome conformational landscape reveals mechanisms of peptidase gating. *Cell Rep.* **24**, 1301–1315 (2018).
13. Wehmer, M. et al. Structural insights into the functional cycle of the ATPase module of the 26S proteasome. *Proc. Natl Acad. Sci. USA* **114**, 1305–1310 (2017).
14. Schweitzer, A. et al. Structure of the human 26S proteasome at a resolution of 3.9 Å. *Proc. Natl Acad. Sci. USA* **113**, 7816–7821 (2016).
15. Lasker, K. et al. Molecular architecture of the 26S proteasome holocomplex determined by an integrative approach. *Proc. Natl Acad. Sci. USA* **109**, 1380–1387 (2012).
16. Lander, G. C. et al. Complete subunit architecture of the proteasome regulatory particle. *Nature* **482**, 186–191 (2012).
17. da Fonseca, P. C., He, J. & Morris, E. P. Molecular model of the human 26S proteasome. *Mol. Cell* **46**, 54–66 (2012).
18. Huang, X., Luan, B., Wu, J. & Shi, Y. An atomic structure of the human 26S proteasome. *Nat. Struct. Mol. Biol.* **23**, 778–785 (2016).
19. Luan, B. et al. Structure of an endogenous yeast 26S proteasome reveals two major conformational states. *Proc. Natl Acad. Sci. USA* **113**, 2642–2647 (2016).
20. Beck, F. et al. Near-atomic resolution structural model of the yeast 26S proteasome. *Proc. Natl Acad. Sci. USA* **109**, 14870–14875 (2012).
21. Śledź, P. et al. Structure of the 26S proteasome with ATP-γS bound provides insights into the mechanism of nucleotide-dependent substrate translocation. *Proc. Natl Acad. Sci. USA* **110**, 7264–7269 (2013).
22. Ding, Z. et al. High-resolution cryo-EM structure of the proteasome in complex with ADP-AlFx. *Cell Res.* **27**, 373-385 (2017).
23. Saeki, Y., Isono, E. & Toh-E, A. Preparation of ubiquitinated substrates by the PY motif-insertion method for monitoring 26S proteasome activity. *Methods Enzymol.* **399**, 215–227 (2005).
24. Verma, R., McDonald, H., Yates, J. R., III & Deshaies, R. J. Selective degradation of ubiquitinated Sic1 by purified 26S proteasome yields active S phase cyclin-Cdk. *Mol. Cell* **8**, 439–448 (2001).
25. Lu, Y., Lee, B. H., King, R. W., Finley, D. & Kirschner, M. W. Substrate degradation by the proteasome: a single-molecule kinetic analysis. *Science* **348**, 1250834 (2015).
26. Galan, J.-M. & Haguenauer-Tsapis, R. Ubiquitin Lys63 is involved in ubiquitination of a yeast plasma membrane protein. *EMBO J.* **16**, 5847-5854 (1997).
27. Whitby, F. G. et al. Structural basis for the activation of 20S proteasomes by 11S regulators. *Nature* **408**, 115-120 (2000).
28. Zhang, F. et al. Structural insights into the regulatory particle of the proteasome from *Methanocaldococcus jannaschii*. *Mol. Cell* **34**, 473–484 (2009).
29. Worden, E. J., Dong, K. C. & Martin, A. An AAA motor-driven mechanical switch in Rpn11 controls deubiquitination at the 26S proteasome. *Mol. Cell* **67**, 799–811. e8 (2017).
30. Pathare, G. R. et al. Crystal structure of the proteasomal deubiquitylation module Rpn8–Rpn11. *Proc. Natl Acad. Sci. USA* **111**, 2984–2989 (2014).
31. Worden, E. J., Padovani, C. & Martin, A. Structure of the Rpn11–Rpn8 dimer reveals mechanisms of substrate deubiquitination during proteasomal degradation. *Nat. Struct. Mol. Biol.* **21**, 220–227 (2014).
32. Dambacher, C. M., Worden, E. J., Herzik, M. A., Martin, A. & Lander, G. C. Atomic structure of the 26S proteasome lid reveals the mechanism of deubiquitinase inhibition. *eLife* **5**, e13027 (2016).
33. Zhang, X. & Wigley, D. B. The 'glutamate switch' provides a link between ATPase activity and ligand binding in AAA+ proteins. *Nat. Struct. Mol. Biol.* **15**, 1223–1227 (2008).
34. He, J. et al. The structure of the 26S proteasome subunit Rpn2 reveals its PC repeat domain as a closed toroid of two concentric α-helical rings. *Structure* **20**, 513–521 (2012).
35. Snoberger, A., Brettrager, E. J. & Smith, D. M. Conformational switching in the coiled-coil domains of a proteasomal ATPase regulates substrate processing. *Nat. Commun.* **9**, 2374 (2018).
36. Glynn, S. E., Martin, A., Nager, A. R., Baker, T. A. & Sauer, R. T. Structures of asymmetric ClpX hexamers reveal nucleotide-dependent motions in a AAA+ protein-unfolding machine. *Cell* **139**, 744–756 (2009).
37. Kim, Y. C., Snoberger, A., Schupp, J. & Smith, D. M. ATP binding to neighbouring subunits and intersubunit allosteric coupling underlie proteasomal ATPase function. *Nat. Commun.* **6**, 8520 (2015).
38. Horwitz, A. A. et al. ATP-induced structural transitions in PAN, the proteasome-regulatory ATPase complex in Archaea. *J. Biol. Chem.* **282**, 22921–22929 (2007).
39. Puchades, C. et al. Structure of the mitochondrial inner membrane AAA+ protease YME1 gives insight into substrate processing. *Science* **358**, eaao0464 (2017).
40. Ripstein, Z. A., Huang, R., Augustyniak, R., Kay, L. E. & Rubinstein, J. L. Structure of a AAA+ unfoldase in the process of unfolding substrate. *eLife* **6**, e25754 (2017).
41. Gates, S. N. et al. Ratchet-like polypeptide translocation mechanism of the AAA+ disaggregase Hsp104. *Science* **357**, 273–279 (2017).
42. Han, H., Monroe, N., Sundquist, W. I., Shen, P. S. & Hill, C. P. The AAA ATPase Vps4 binds ESCRT-III substrates through a repeating array of dipeptide-binding pockets. *eLife* **6**, e31324 (2017).
43. Deville, C. et al. Structural pathway of regulated substrate transfer and threading through an Hsp100 disaggregase. *Sci. Adv.* **3**, e1701726 (2017).
44. Monroe, N., Han, H., Shen, P. S., Sundquist, W. I. & Hill, C. P. Structural basis of protein translocation by the Vps4-Vta1 AAA ATPase. *eLife* **6**, e24487 (2017).
45. Enemark, E. J. & Joshua-Tor, L. Mechanism of DNA translocation in a replicative hexameric helicase. *Nature* **442**, 270–275 (2006).
46. Thomsen, N. D. & Berger, J. M. Running in reverse: the structural basis for translocation polarity in hexameric helicases. *Cell* **139**, 523–534 (2009).

Acknowledgements We thank Y. Saeki for the plasmids expressing Sic1^{PY} and WW-HECT; H. Huang for proteasome-expressing cell lines; D. Yu, J. Xu, Y. Ma, C. Fan and J. Jackson for technical support; and S. Elsasser for critical reading of the manuscript. This work was funded in part by an Intel Corporation academic grant, the Thousand Talents Plan of China, National Natural Science Foundation of China grant nos. 11774012 and 91530321, the Peking-Tsinghua Center for Life Sciences (Y.M.), NIH grant GM43601 (D.F.) and an Edward Mallinckrodt, Jr. Foundation award (Y.L.). The cryo-EM data were collected from the Electron Microscopy Laboratory and Cryo-EM Platform at Peking University. Initial cryo-EM screening was performed in part at the Center for Nanoscale Systems at Harvard University supported by the National Science Foundation under NSF award no. 1541959 and NIH grant AI100645. Data processing was performed in part in the Sullivan cluster, which is funded in part by a gift from Mr. and Mrs. Daniel J. Sullivan, Jr., and in the Weiming No.1 and Life Science No. 1 High-Performance Computing Platform at Peking University.

Reviewer information *Nature* thanks C. Hill and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Y.D. purified proteins, conducted biochemical analysis and prepared samples for imaging. Y.D., S.Z., Z.W., X.L., W.L.W., Y.Z. and S.S.-M. collected data. S.Z. and Z.W. processed data and refined the maps. Y.D., S.Z., Y.L. and D.F. contributed to structural analysis and manuscript preparation. Y.M. conceived and supervised this study, devised the methods, performed atomic modelling, analysed the structures and wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0736-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0736-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Y.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Preparation of polyubiquitylated Sic1^{PY}. Sic1^{PY} and WW-HECT, which consists of both WW and HECT domains of Rsp5, were chosen as the model substrate and the E3 ubiquitin ligase, respectively[23]. The PY motif is recognized by WW domains in the Rsp5 family of ligases. In the Sic1^{PY} construct used in this study, a PY motif (Pro-Pro-Pro-Ser) was inserted into the N terminus (MTPSTPPSRGTRYLA) of the Cdk inhibitor Sic1[7,24], resulting in a modified N terminus of MTPSTPPPPPSSRGTRYLA, in which the first residue to be ubiquitylated is likely to be Lys18 of Sic1^{PY}. WW-HECT was derived from wild-type Rsp5 of *Saccharomyces cerevisiae* by deletion of its N-terminal 220 amino acids, and conjugates ubiquitin chains by Lys63 linkage[23,26]. Both proteins were expressed in *E. coli* and purified as previously reported[23]. Plasmids expressing Sic1^{PY} and WW-HECT were gifts from Y. Saeki (Tokyo Metropolitan Institute of Medical Science). After plasmid transformation into BL21 (DE3) cells, cultures were grown to an optical density (OD_{600}) of 0.7 in LB medium with 50 μg/ml ampicillin. Cultures were cooled to 30 °C, isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to 0.5 mM and incubation proceeded for 3 h. After being collected by centrifugation (3,000*g*, 10 min), cells were suspended and lysed by sonication in 50 mM PBS (pH 7.0) containing 300 mM NaCl, 10% glycerol, 1 mM DTT, 0.2% Triton X-100 and 1× protease inhibitor cocktail. The supernatant was recovered after centrifugation (15,000*g*, 30 min), then incubated with pre-equilibrated TALON resin for 2 h at 4 °C. After this binding step, the resin was washed with 20 column volumes of 50 mM Tris-HCl (pH 7.5) containing 100 mM NaCl, 10% glycerol, and 1 mM DTT. Sic1^{PY} was then eluted with the same buffer containing 150 mM imidazole. The eluted sample was further purified by fast protein liquid chromatography (FPLC; Superdex 75; 0.25 ml/min), using a buffer containing 50 mM Tris-HCl (pH 7.5), 100 mM NaCl, and 10% glycerol.

To purify WW-HECT, plasmid-transformed BL21 (DE3) cells were grown to an OD_{600} of 0.5 in LB medium with 50 μg/ml ampicillin. The culture was then cooled to 20 °C and WW-HECT synthesis induced by the addition of IPTG to 0.2 mM. Cells were collected 15 h after induction and lysed through the same procedure as used for the Sic1^{PY}. A 15,000*g* supernatant was incubated with pre-equilibrated glutathione sepharose resin for 2 h at 4 °C. The resin was washed with 20 column volumes of washing buffer (50 mM Tris-HCl (pH 7.5) containing 100 mM NaCl, 10% glycerol and 1 mM DTT), then incubated with the same buffer containing PreScission protease for 12 h at 4 °C. The resin was removed by centrifugation and the supernatant was then applied to FPLC (Superdex 75) as described above.

To ubiquitinate Sic1^{PY}, 40 μg/ml Sic1^{PY}, 500 nM UBE1 (Boston Biochem), 2 μM UBCH5A (Boston Biochem), 100 μg/ml WW-HECT and 1 mg/ml ubiquitin (Boston Biochem) were incubated in reaction buffer (50 mM Tris-HCl [pH 7.5],100 mM NaCl, and 10% glycerol, 2 mM ATP, 10 mM $MgCl_2$, and 1 mM DTT) for 3 h at room temperature. Pre-equilibrated TALON resin was then incubated with the this sample for 1 h at 4 °C. After the resin was washed with 20 column volumes of the wash buffer (50 mM Tris-HCl (pH 7.5), 100 mM NaCl, 10% glycerol), the polyubiquitinated Sic1^{PY} (PUb-Sic1^{PY}) was eluted with the same buffer containing 150 mM imidazole. The elution was applied to an Amicon ultrafiltration device with 30K molecular cut-off for removal of imidazole. The ubiquitination reaction was examined by western blotting with anti-T7 antibody (Extended Data Fig. 1f).

Purification of the human 26S proteasome. Human proteasomes were affinity-purified on a large scale from a stable HEK293 cell line containing HTBH (hexahistidine, TEV cleavage site, biotin, and hexahistidine)-tagged RPN11 (a gift from L. Huang, University of California, Irvine)[47]. The cells were Dounce-homogenized in a lysis buffer (50 mM PBS (pH 7.5), 10% glycerol, 5 mM $MgCl_2$, 0.5% NP-40, 5 mM ATP and 1 mM DTT) containing protease inhibitors. Lysates were cleared by centrifugation (20,000*g*, 30 min), then incubated with NeutrAvidin agarose resin (Thermo Scientific) for 3 h at 4 °C. The beads were washed with excess lysis buffer followed by wash buffer (50 mM Tris-HCl (pH 7.5), 1 mM $MgCl_2$ and 1 mM ATP). 26S proteasomes were cleaved from the beads using TEV protease (Invitrogen). The resin was removed by centrifugation and the supernatant was then further purified by gel filtration on a Superose 6 10/300 GL column at a flow rate of 0.15 ml/min in buffer (30 mM Hepes (pH 7.5), 60 mM NaCl, 1 mM $MgCl_2$, 10% glycerol, 0.5 mM DTT, 0.6 mM ATP). Gel-filtration fractions were concentrated to about 2 mg/ml and the buffer was exchanged to 50 mM Tris-HCl (pH 7.5), 100 mM NaCl, 1 mM ATP and 10% glycerol (Extended Data Fig. 1a–c).

Biochemical verification of the substrate-bound human proteasome. To verify the preparation of the polyubiquitylated (PUb)-Sic1^{PY}, we performed degradation assays on the PUb-Sic1^{PY} using our purified human 26S proteasome. We incubated 100 nM proteasome with 20 μg/ml PUb-Sic1^{PY} for 10 min at 37 °C in a buffer containing 50 mM Tris-HCl (pH 7.5), 100 mM NaCl, 10% glycerol and 5 mM ATP. The reaction was stopped by adding SDS loading buffer and 100 mM DTT. Samples were collected at 2, 5, and 10 min. The Sic1^{PY} degradation reaction was followed by western blotting with the anti-T7 antibody (Extended Data Fig. 1h).

To verify the formation of a proteasome–substrate complex in our cryo-EM imaging experiments, we crosslinked the proteasome–substrate complexes and examined them by native gel electrophoresis. However, crosslinking was not used for the sample preparation for cryo-EM data collection, to preserve the native states of substrate interactions with the proteasome. Before crosslinking, proteasome and PUb-Sic1^{PY} samples were first exchanged to a buffer containing 50 mM PBS (pH 7.5), 100 mM NaCl, 10% glycerol, 1 mM ATP using Zeba Micro Spin Desalting Columns (7K, Thermo Fisher). Then, 1 μl of 2 mg/ml PUb-Sic1 and 1 μl of 1 mg/ml proteasome were mixed with 17 μl of the same buffer for 30 s. Then, 1 mM ATPγS was immediately added, after which 1 μl of freshly prepared 2.3% solution of glutaraldehyde was added and incubated for 15 min at 37 °C. The crosslinked complex was then examined by native gel electrophoresis (Extended Data Fig. 1g).

Cryo-EM imaging and data collection. We incubated 10 μl of 2 mg/ml proteasome with 9 μl of 2 mg/ml PUb-Sic1^{PY} (molar ratio ~3:1 for substrate:proteasome) for 30 s (50 mM Tris-HCl (pH 7.5), 100 mM NaCl, 10% glycerol and 1 mM ATP) at room temperature and 1 μl 20 mM ATPγS was then immediately added to the solution. To remove the glycerol, the complex system was applied to Zeba Micro Spin Desalting Columns (7K, Thermo Fisher), exchanging the buffer to 50 mM Tris-HCl (pH 7.5) containing 100 mM NaCl, 1mM ATP and 1 mM ATPγS. The glycerol removal process usually took about 10 min before cryo-plunging. We added 0.005% NP-40 to the proteasome solution immediately before cryo-plunging. Cryo-EM grids were prepared with FEI Vitrobot Mark IV. C-flat grids (R1/1 and R1.2/1.3, 400 Mesh, Protochips) were glow-discharged before a 2.5-μl drop of 1.5 mg/ml substrate-engaged proteasome solution was applied to the grids in an environmentally controlled chamber with 100% humidity and temperature fixed at 4 °C. After 2 s of blotting, the grid was plunged into liquid ethane and then transferred to liquid nitrogen. The cryo-grids were initially screened at a nominal magnification of 235,000× in an FEI Tecnai Arctica microscope, equipped with an Autoloader and an acceleration voltage of 200 kV. Good-quality grids were transferred to an FEI Titan Krios G2 microscope equipped with the post-column Gatan BioQuantum energy filter connected to Gatan K2 Summit direct electron detector. Coma-free alignment was manually optimized and parallel illumination was verified before data collection. Cryo-EM data were collected semi-automatically by Leginon[48] version 3.1 and SerialEM[49] with the Gatan K2 Summit operating (Gatan) in a super-resolution counting mode and with the Gatan BioQuantum operating in the zero-loss imaging mode (10-μm energy slit). A total exposure time of 10 s with 250 ms per frame resulted in a 40-frame movie per exposure with an accumulated dose of 44 electrons/Å^2. The calibrated physical pixel size and the super-resolution pixel size are 1.37 Å and 0.685 Å, respectively. The raw data were saved at the pixel size of 0.685 Å. The defocus in data collection was set in the range of −0.7 to −3.0 μm. A total of 44,664 movies was collected throughout eight sessions of data collection.

Cryo-EM data processing and reconstruction. The micrograph frames of 44,664 raw movies were aligned and averaged with the MotionCor2 program[50] at a super-resolution pixel size of 0.685 Å. Each drift-corrected micrograph was used for the determination of the micrograph CTF parameters with program Gctf[51]. We picked 2,669,687 particles of the 26S proteasome using the program deepEM[52]. Reference-free 2D classification and 3D classification were carried out with two-fold binned data with a pixel size of 1.37 Å in both RELION 2.1[53] and ROME, which combined maximum-likelihood based image alignment and statistical machine-learning based classification[54]. Focused 3D classification, which we used in the later stage of data processing, and high-resolution refinement were mainly done with RELION 2.1. Map reconstruction and local resolution calculation were finished with programs in both RELION 2.1 and ROME. A substantial part of the data processing, mostly 2D/3D classification, was performed with a 1024-core CPU cluster equipped with 64 Intel Xeon Gold 6142 (2.6 GHz 16-core) CPUs, a NVIDIA DGX-1 supercomputing system equipped with 8 Tesla V100 GPUs or a 10-node GPU cluster equipped with 40 Tesla V100 GPUs.

We applied a hierarchical 3D classification strategy to analyse the very large dataset (Extended Data Fig. 2). The entire data-processing procedure consisted of four steps. In the first step, we separated doubly capped proteasome particles from singly capped ones through several rounds of 2D and 3D classification. This resulted in 1,552,828 doubly capped particles and 478,919 singly capped ones. These particles were aligned to the consensus models of doubly and singly capped proteasomes to obtain their approximate shift and angular parameters. With these parameters, each complete doubly capped particle was split into two pseudo-singly capped particles by re-centring the box onto the RP–CP subcomplex. Then the box size of pseudo-singly capped particles and true singly capped particles was shrunk to 600 × 600. This is an effective way to reduce irrelevant heterogeneity owing to conformational variations, and to improve map resolution[11,14]. There were 3,584,040 particles in the dataset chosen for the following steps of analysis. In the second step, we focused on the gate of the CP. Several rounds of 2D and 3D

© 2019 Springer Nature Limited. All rights reserved.

classification were done to distinguish the states of the CP gate—that is, separating the S_A-like closed-gate states from open-gate or S_D-like states. It was obvious that the RP subcomplex of the S_D-like states rotates by a large angle compared to the S_A-like states[10,11]. The RP–CP subcomplex was masked during the 3D classification. There were 732,666 particles in S_A-like states and 2,521,686 particles in S_D-like states after this step[10,11]. In the third step, we used focused 3D classification[53] to further classify within these two different states. As the CP is structurally stable, we did refinement with the CP masked, so that we could determine the *x*–*y* shift and angular parameters of all particles when they were aligned against the reference of the CP. Using these parameters, we continued 3D classification with the RP masked and with alignment skipped. This means we classified images based only on structural changes in the RP relative to the CP. After the classification, we clearly saw that the RP markedly swings and rotates against the CP. By focusing on the variation of substrate interactions with the AAA-ATPase and RPN10/11, we classified these particles into 5 major states, designated E_A, E_B, E_C, E_{D1} and E_{D2}, respectively, accounting for 7.8%, 14.8%, 9.9%, 27.2%, and 40.1% of the particles. In the final step, we used focused classification to further detect substantial structural changes within each of these five states. After auto-refinement with the RP masked, we continued skip-alignment classification with the lid, the AAA-ATPase or certain combinations of RP subunits masked. Application of differential masks depended on specific structural characteristics of different states. For example, RPN1 in state E_B was partially blurred without further 3D classification. We therefore masked RPN1 together with ATPase in 3D classification, which resulted in improvement of its density quality in certain 3D classes. The whole RP complex is highly dynamic in state E_C. Thus, we performed further classification with the whole RP masked, resulting in two distinct states named E_{C1} and E_{C2}. E_{C1} showed a clear ubiquitin density, which is absent in E_{C2}. Similarly, we also obtained an intermediate state from initial E_B dataset, named E_{A2}, which showed a ubiquitin-binding mode different from that in E_{A1}.

The final refinement of each state was done using data with a pixel size of 1.37 Å that were binned by two-fold from the raw data in the super-counting mode. Based on the in-plane shift and Euler angle of each particle from the last iteration of refinement, we reconstructed the two half-maps of each state using raw single-particle images at the super-counting mode with a pixel size of 0.685 Å. To enhance the local density quality for each state, we applied two types of local mask in the last several iterations of refinement, one focusing on the complete RP and the other focusing on the CP and ATPase components, which yielded two maps for each state that showed improved local resolution in the lid and CP, respectively. For each state, the maps refined by differential masking were merged in Fourier space into a single map. This procedure was also applied for the half maps before Fourier shell correlation (FSC) calculation. Because states E_{A1} and E_{A2} exhibit identical structures in their CP and AAA-ATPase components, we combined them together and refined the combined dataset by applying the CP–ATPase mask. The final reconstructions of the combined E_A, E_{A1}, E_{A2}, E_B, E_{C1}, E_{C2}, E_{D1} and E_{D2} datasets used CTF parameters calculated at the level of individual particles with Gctf and gave overall resolutions of 2.8 Å, 3.0 Å, 3.2 Å, 3.3 Å, 3.5 Å, 3.6 Å, 3.3 Å and 3.2 Å, respectively, measured by the gold-standard FSC at 0.143-cutoff on two separately refined and merged half maps. Prior to visualization, all density maps were sharpened by applying a negative *B*-factor. Local resolution variations were further estimated using ResMap on the two half maps refined independently[55].

Atomic model building and refinement. The higher-resolution cryo-EM maps allowed us to refine atomic models with improved quality and to extend sequence register beyond the published structures of the substrate-free proteasomes through de novo modelling (Extended Data Table 1). Given that we did not stall the substrates in a homogeneous location during their degradation, and also that substrate translocation through the proteasome is not sequence-specific, the substrate densities were modelled using polypeptide chains without assignment of amino acid sequence (except for the lysine residue forming a visible isopeptide bond with ubiquitin in state E_B).

To build the initial atomic model of the substrate-bound 26S proteasome complex, we used previously published human proteasome structures[11] as starting models and rebuilt each atomic model in Coot[56] for each of the seven conformational states. In states E_{D1} and E_{D2}, many residues at the N terminus of the CC domain of RPT1 and RPT2 and the C-terminal toroidal domain of RPN2 that were missing in other states and in the previously published substrate-free structures[10–22] were shown as reliable densities with flanking of large side chains. These high-resolution features allowed us to conduct de novo tracing of these previously missing elements, including a newly identified helix of RPN2 residing in a long loop (residues 820–871) emanating from the RPN2 toroidal domain (Extended Data Fig. 9). In all previously published cryo-EM structures of human 26S proteasomes, the local resolution of the lid subcomplex was generally worse than 4.9 Å and was insufficient to ensure the correct register of the side chains. Our density maps of all states—particularly E_B and $E_{D1,2}$—exhibit substantially improved local resolution in the lid subcomplex (Extended Data Table 1, Extended Data Fig. 3), allowing us to rebuild the majority of the lid subcomplex. The local resolution of RPN1 in states E_B and $E_{D1,2}$ also reached 4–5 Å, allowing us to improve the backbone model and make partial side-chain registers. RPN1 has very poor local resolution in states E_{C1} and E_{C2} precluding de novo atomic modelling. Thus, we used the improved atomic model of RPN1 from states E_B and $E_{D1,2}$ to fit the poor RPN1 densities of $E_{C1,2}$ as a rigid body.

The nucleotide densities are of sufficient quality for differentiating ADP from ATP, which allowed us to build the atomic models of ADP and ATP into their densities. A resolution of no worse than 3.6 Å may be required to distinguish ADP from ATP, because ATP adds an extra size of 2.46 Å with its γ-phosphate and three additional oxygen atoms relative to ADP. The magnesium ion bound to ATP was well-resolved in all states except E_{C2} (Extended Data Fig. 3c). By contrast, no magnesium-ion density was observed around the ADP-assigned nucleotide density except for ADP in RPT5 of E_A and in RPT3 of E_{D2}. Except for states E_{A1} and E_{A2}, at least one of the ATPases in each state has a very poor nucleotide density quality in its nucleotide-binding site (Extended Data Fig. 7). Although there are visible extra densities in the nucleotide-binding site at a low contour level in most of the apo-like ATPase subunits after the protein structures are in place, these weak extra densities are insufficient for even fitting a complete ADP with good confidence. For instance, some of them may allow fitting of ribose and/or α-phosphate—but then β-phosphate is totally out of density. To avoid over-interpretation and to practice prudence in high-quality atomic modelling, we avoided building atomic models of nucleotides into these poor densities at all and referred to the corresponding ATPases as the 'apo-like state' throughout this study. The poor extra densities in the nucleotide-binding sites of these apo-like ATPases are likely to reflect partial or low occupancy or unstable binding of nucleotide, which is expected when the nucleotide-binding site undergoes nucleotide exchange.

Atomic model refinement was conducted in Phenix[57] with its real-space refinement program. We used both simulated annealing and global minimization with NCS, rotamer and Ramachandran constraints. Partial rebuilding, model correction and density-fitting improvement in Coot[56] were iterated after each round of atomic model refinement in Phenix[57]. The improved atomic models were then refined again in Phenix, followed by rebuilding in Coot[56]. The refinement and rebuilding cycle was repeated until the model quality reached expectation (Extended Data Table 1).

Structural analysis and visualization. All figures of structures were plotted in Chimera[58], PyMOL[59], or Coot[56]. Structural alignment and comparison were performed in both PyMOL and Chimera. Interaction analysis between adjacent subunits was performed using PISA[60].

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

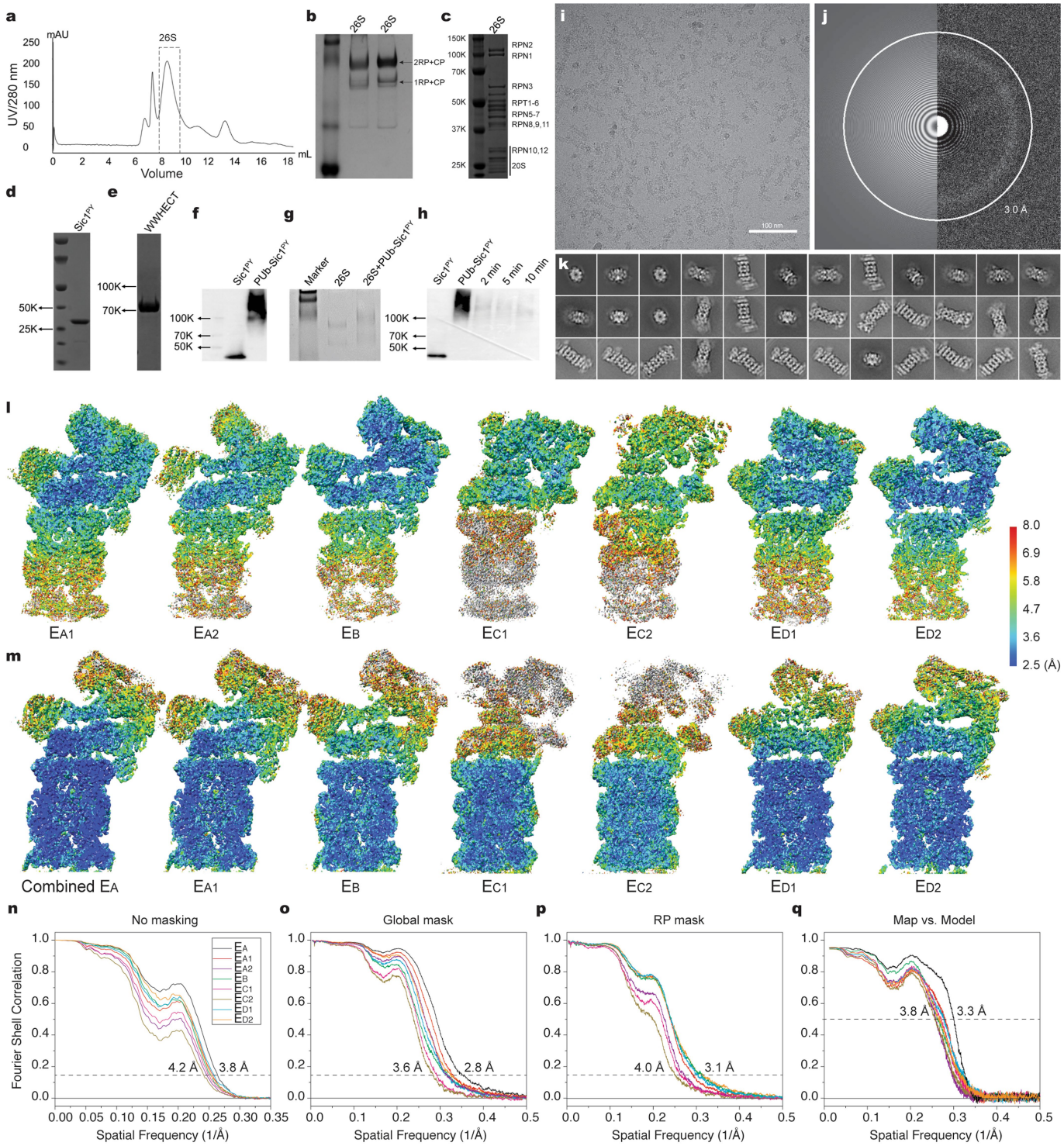
Cryo-EM maps have been deposited in the Electron Microscopy Data Bank (EMDB) under accession codes EMD-9215 (the combined E_A refined with CP-ATPase mask), EMD-9216 (whole E_{A1}), EMD-9217 (whole E_{A2}), EMD-9218 (whole E_B), EMD-9219 (whole E_{C1}), EMD-9220 (whole E_{C2}), EMD-9221 (whole E_{D1}), EMD-9222 (whole E_{D2}), EMD-9223 (RP of E_{A1}), EMD-9224 (RP of E_{A2}), EMD-9225 (RP of E_B), EMD-9226 (RP of E_{C1}), EMD-9227 (RP of E_{C2}), EMD-9228 (RP of E_{D1}) and EMD-9229 (RP of E_{D2}). Each EMDB entry includes three maps: (1) the low-pass-filtered map without amplitude correction as a default; (2) the low-pass-filtered map with amplitude correction by a negative *B*-factor shown in Extended Data Table 1; and (3) the raw map without any post-processing such as low-pass-filtering and amplitude correction. Coordinates are available from the RCSB Protein Data Bank under accession codes 6MSB (whole E_{A1}), 6MSD (whole E_{A2}), 6MSE (whole E_B), 6MSG (whole E_{C1}), 6MSH (whole E_{C2}), 6MSJ (whole E_{D1}) and 6MSK (whole E_{D2}). Raw data are available from the corresponding author upon request.

47. Wang, X. et al. Mass spectrometric characterization of the affinity-purified human 26S proteasome complex. *Biochemistry* **46**, 3553–3565 (2007).
48. Suloway, C. et al. Automated molecular microscopy: the new Leginon system. *J. Struct. Biol.* **151**, 41–60 (2005).
49. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
50. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
51. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
52. Zhu, Y., Ouyang, Q. & Mao, Y. A deep convolutional neural network approach to single-particle recognition in cryo-electron microscopy. *BMC Bioinformatics* **18**, 348 (2017).
53. Kimanius, D., Forsberg, B. O., Scheres, S. H. & Lindahl, E. Accelerated cryo-EM structure determination with parallelisation using GPUs in RELION-2. *eLife* **5**, e18722 (2016).

© 2019 Springer Nature Limited. All rights reserved.

54. Wu, J. et al. Massively parallel unsupervised single-particle cryo-EM data clustering via statistical manifold learning. *PLoS ONE* **12**, e0182130 (2017).
55. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
56. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
57. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
58. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
59. The PyMOL Molecular Graphics System Version 1.8 (Schrödinger, LLC).
60. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).

© 2019 Springer Nature Limited. All rights reserved.

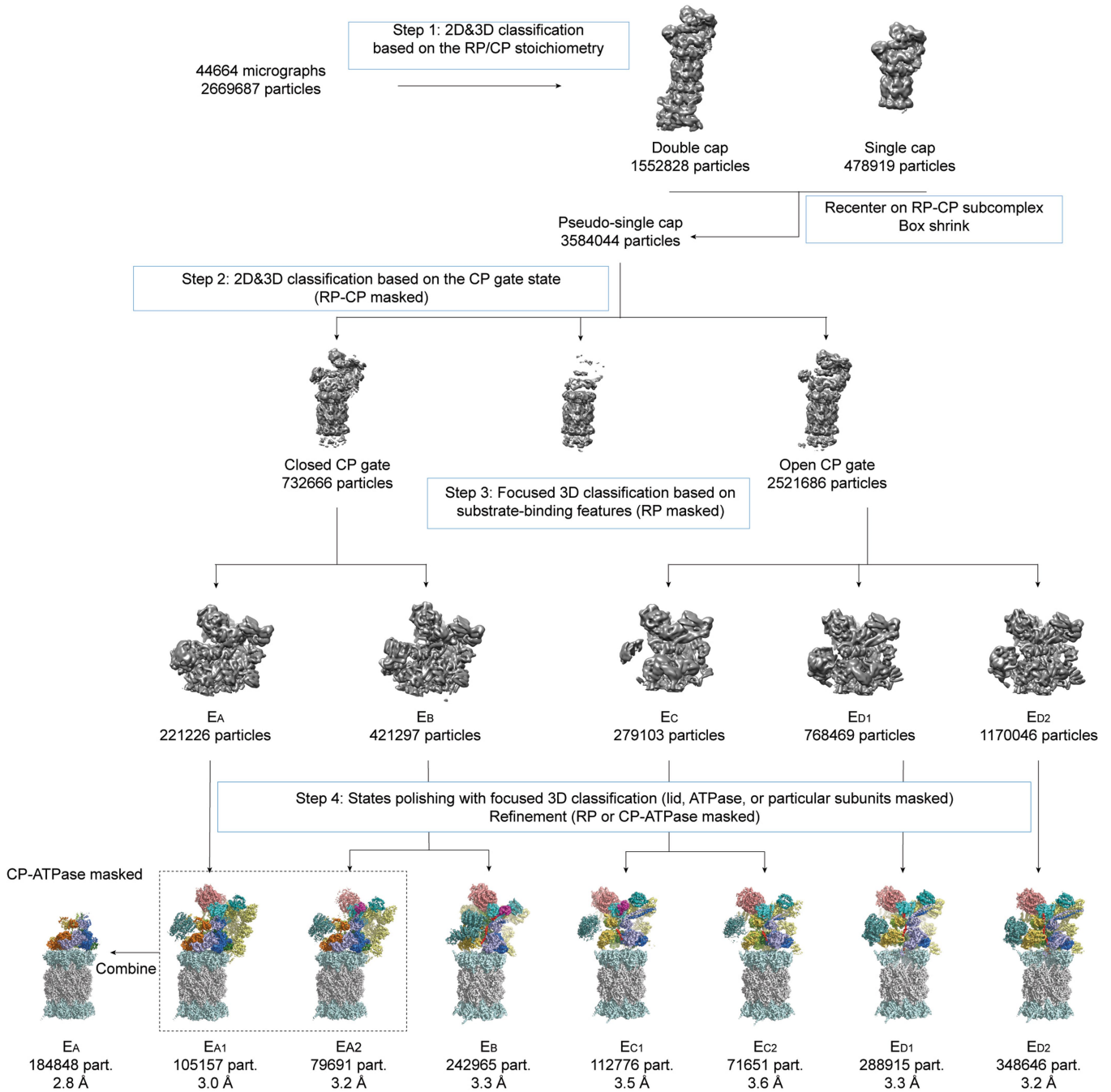


Extended Data Fig. 1 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

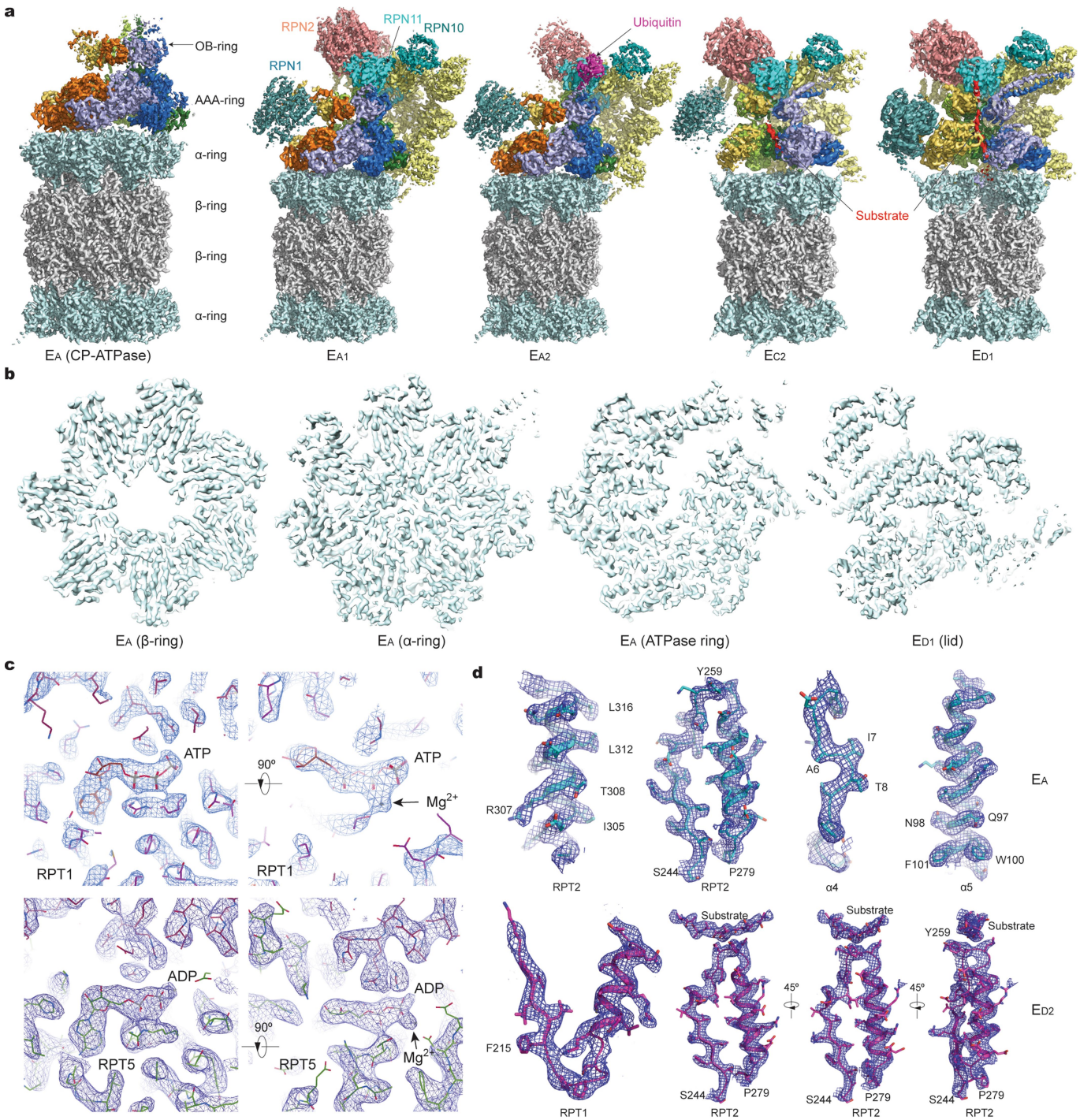
Extended Data Fig. 1 | Characterization and structure determination of the substrate-engaged human proteasome. a, FPLC purification of the human 26S proteasome on Superose 6 10/300 GL. The dashed box shows the fraction taken for structural analysis. **b**, Native PAGE analysis of proteasome purified as in **a**. **c**, SDS–PAGE analysis of proteasome purified as in **a**. **d**, SDS–PAGE analysis of $Sic1^{PY}$ purified through a Superdex 75 column. **e**, SDS–PAGE analysis of WW-HECT purified through Superdex 75. **b**–**e**, Stained with Coomassie blue. **f**, SDS–PAGE/western blot analysis of polyubiquitinated $Sic1^{PY}$. After ubiquitination, the samples were applied to an SDS–polyacrylamide gel, followed by western blotting with anti-T7 antibody. The result suggests that almost all $Sic1^{PY}$ is ubiquitinated. **g**, Native PAGE analysis of proteasomes crosslinked to PUb-$Sic1^{PY}$. Left, no-substrate control. With the addition of the PUb-$Sic1^{PY}$, the proteasome ran slower than without PUb-$Sic1^{PY}$, indicating that the substrate had been captured by the proteasome but not totally degraded when samples were prepared for cryo-EM analysis. **h**, SDS–PAGE/western blot analysis of the degradation of PUb-$Sic1^{PY}$, which was visualized with anti-T7 antibody. This experiment confirms that PUb-$Sic1^{PY}$ is readily degraded by our proteasome samples. **i**, Typical cryo-EM micrograph of the substrate-engaged human proteasome after motion correction. All experiments in **a**–**i** were repeated independently at least three times with similar results. **j**, Power spectrum evaluation of the micrograph shown in **i**. **k**, Gallery of unsupervised class averages calculated by ROME[54] using machine-learning-based clustering. **l**, Local resolution estimation calculated by ResMap[55] on seven maps refined by focusing the mask on the RP component. **m**, Local resolution estimation on seven maps refined by focusing the mask on the CP and ATPase components. **n**, Gold-standard FSC plots of eight maps calculated without masking the separately refined half-maps. **o**, Gold-standard FSC plots of the maps refined by focusing on the CP and ATPase components, calculated with masking of the raw half-maps. **p**, Gold-standard FSC plots of the maps refined by focusing on the RP subcomplex, calculated with masking of the raw half-maps. **q**, Model-map FSC plots calculated by Phenix[57] between each refined map and its corresponding atomic model. For each state, the maps refined by differential masking were merged in Fourier space into a single map, which was used for the model–map FSC calculation. The same colour code is used in **n**–**q**.

© 2019 Springer Nature Limited. All rights reserved.



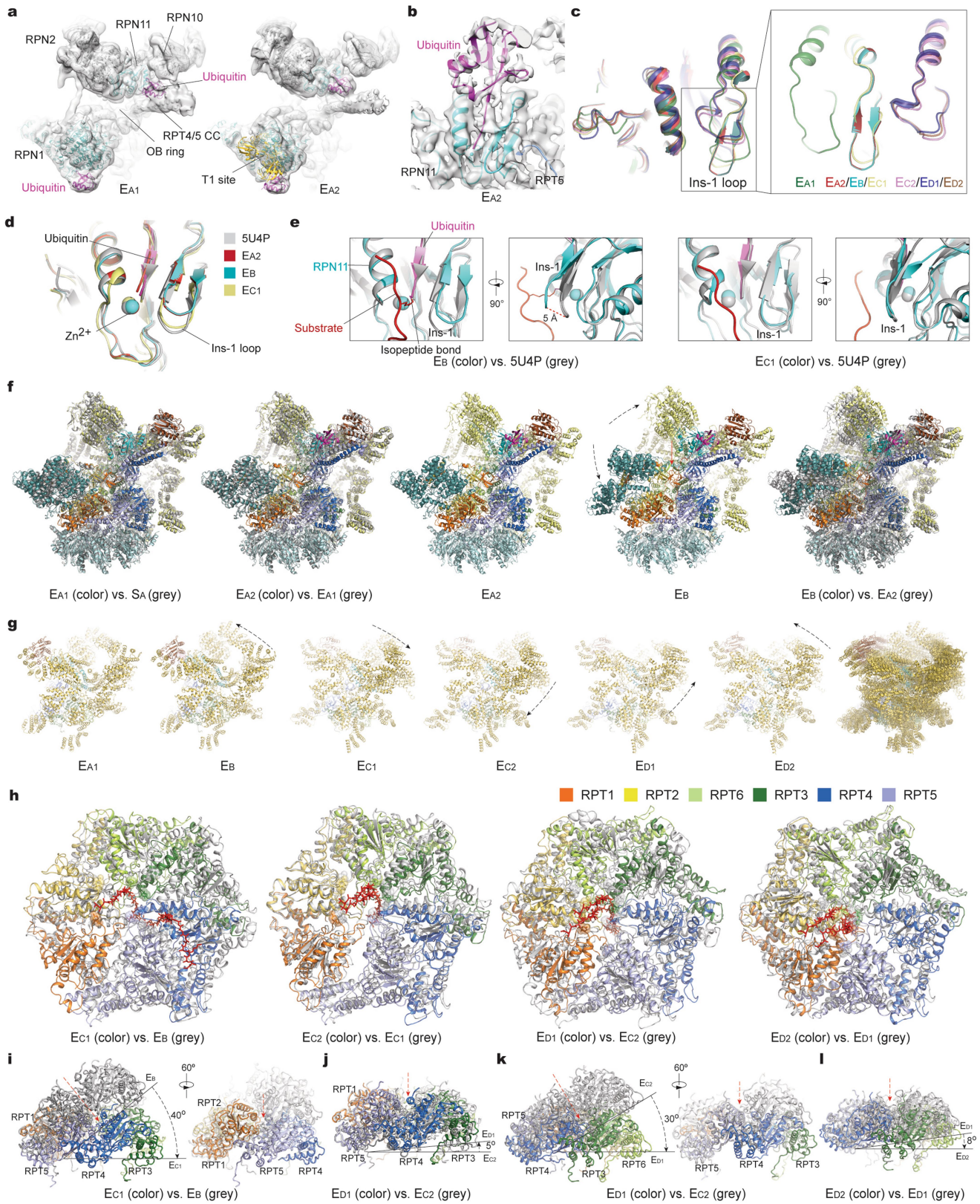
Extended Data Fig. 2 | Focused classification to separate the seven conformational states. The diagram illustrates the four major steps of our hierarchical focused classification strategy. Further detailed iterations of classification in each step are omitted for clarity.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | Cryo-EM maps and quality assessment. a, The five refined cryo-EM maps that are not shown in the main figures. **b**, Typical central cross-sections of the density maps for each of the four subcomplexes (the α-ring, the β-ring and the ATPase ring in state E_A, and the lid in state E_{D1}). **c**, Typical nucleotide-binding pocket densities from the E_{A1} map. The ATP density is compared with the ADP density in two orthogonal perspectives. The magnesium density next to the nucleotide is labelled. **d**, Typical densities of secondary structures and substrate in the proteasome superimposed with their atomic models.

© 2019 Springer Nature Limited. All rights reserved.

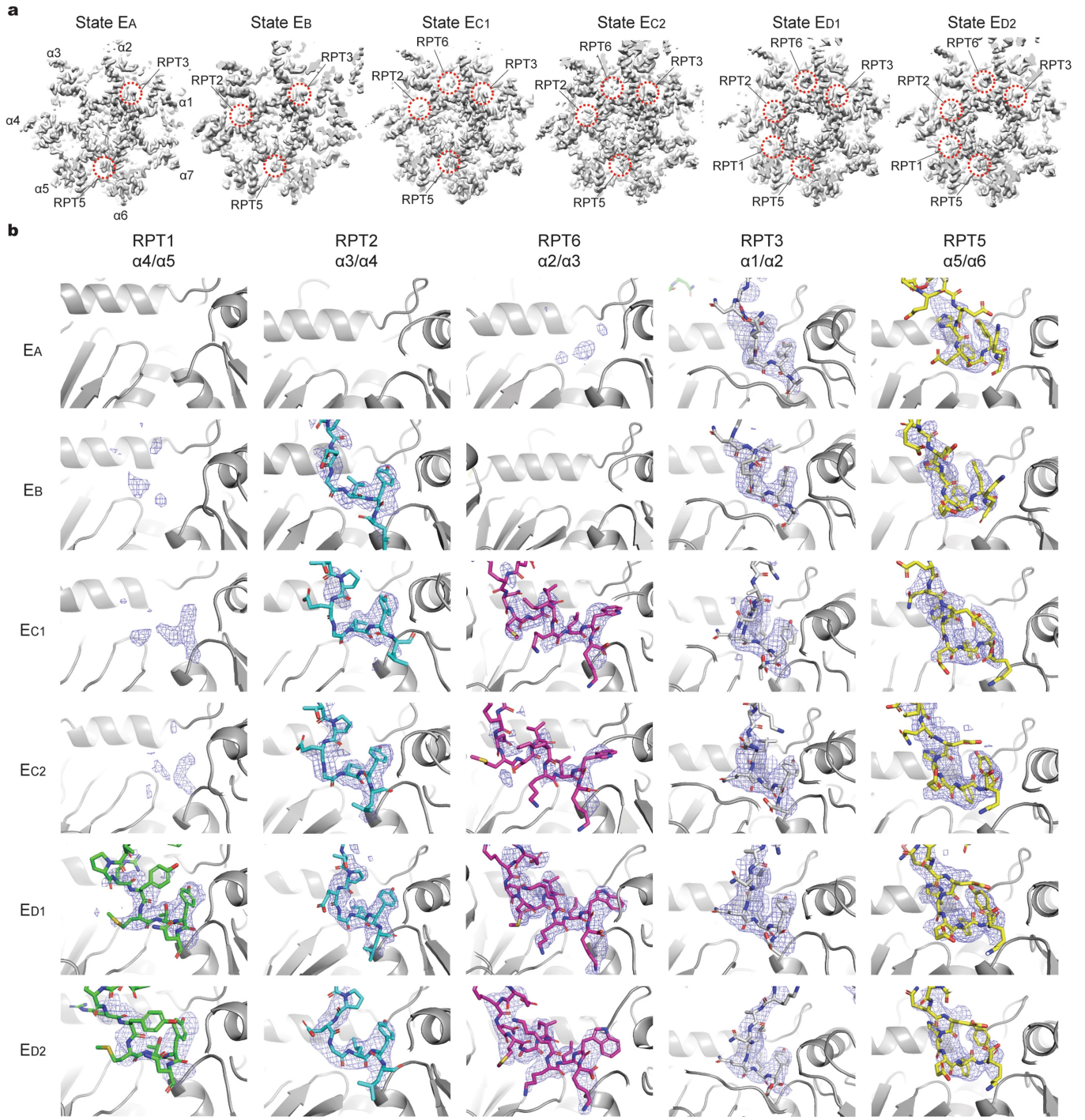


Extended Data Fig. 4 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

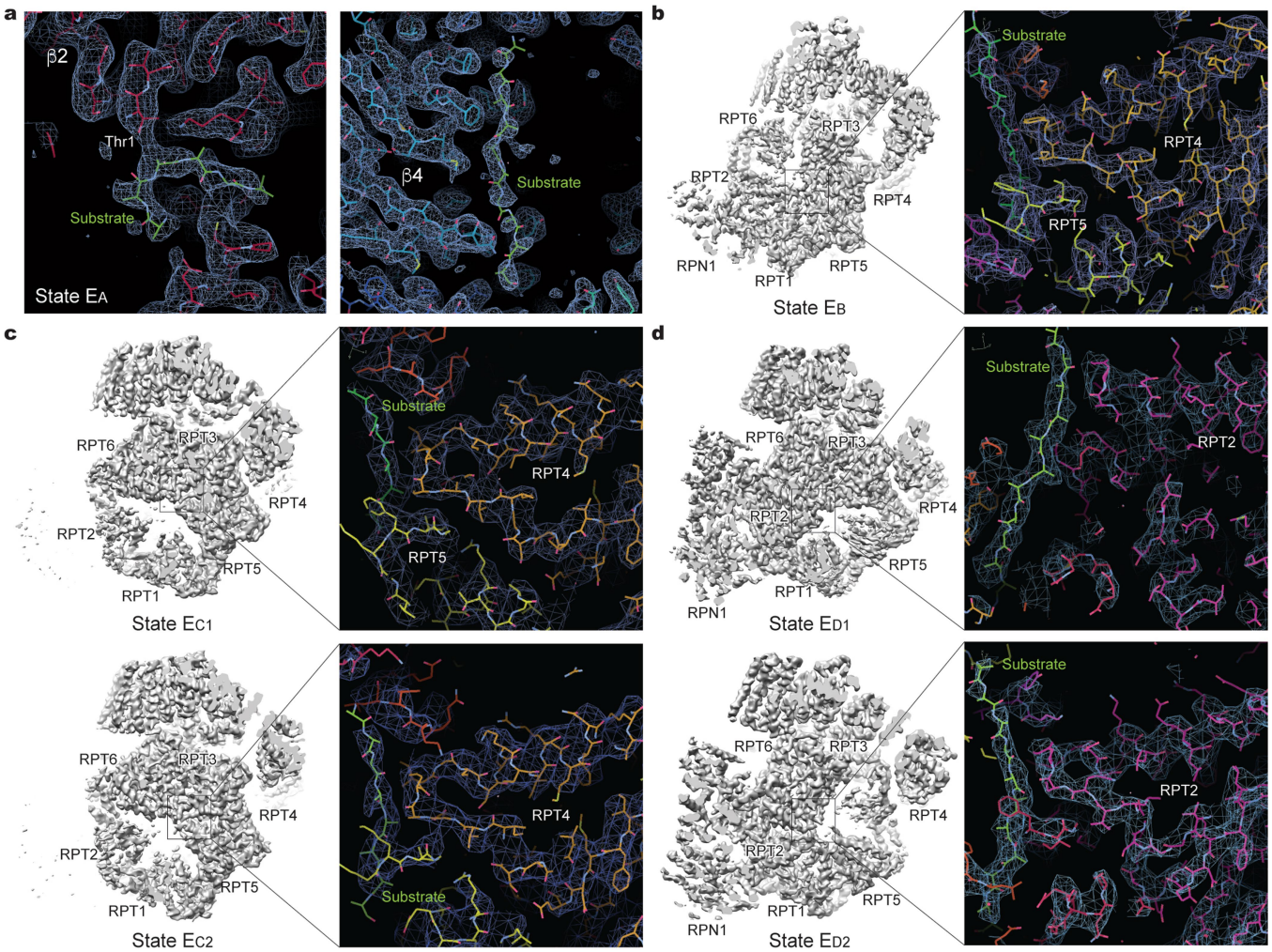
Extended Data Fig. 4 | Key structural features that differentiate the seven conformational states. a, Ubiquitin densities in states E_{A1} (left) and E_{A2} (right). The T1 site is labelled by fitting the yellow cartoon representation of the NMR structure (RCSB Protein Data Bank (PDB) ID 2N3U) of the yeast Rpn1 T1 element in complex with two ubiquitins into our density, showing that the ubiquitin on human RPN1 is bound to a site very close to the yeast Rpn1 T1 site[6]. The density maps are low-pass-filtered to 8 Å to show the ubiquitin features clearly, owing to the lower local resolution of the ubiquitin density in these maps. **b**, The ubiquitin–RPN11–RPT5 interface observed at high resolution in state E_B is also observed in state E_{A2}, albeit at lower local resolution. The E_{A2} density is shown as a transparent surface. **c**, Comparison of the Ins1 loop of RPN11 in different states. **d**, Comparison of the RPN11 structures in states E_{A2}, E_B and E_{C1} around the zinc-binding site and Ins1 region with that in the crystal structure (PDB ID: 5U4P) of a ubiquitin-bound Rpn11–Rpn8 complex from yeast[29]. **e**, Close-up comparison of the RPN11 Ins1 structure between state E_B and 5U4P (left two panels) and between state E_{C1} and 5U4P (right two panels) in two orthogonal perspectives, showing a 5 Å displacement of the Ins1 β-hairpin in E_B relative to 5U4P or E_{C1}. This displacement is not observed between E_{C1} and 5U4P, suggesting that the Ins1 β-hairpin tilt in E_B is mostly to optimize the coordination of the isopeptide bond with the zinc ion. **f**, Comparison of the RP structures of E_A and E_B. **g**, Comparison of lid subcomplex conformations among all states. **h**, Comparison of ATPase ring structures between two successive states. The structures are aligned together against their CP in **f**–**h**. **i**, Side views of the structural comparison of the AAA ring between E_{C1} (colour) and E_B (grey). The large AAA subdomain of RPT1 was used to align the two AAA-ring structures together. A 40° out-of-plane rotation of the large AAA subdomain of RPT1 relative to the AAA ring is observed during disengagement of RPT1–RPT2 from the substrate. The right panel, rotated vertically against the left panel, shows that the out-of-plane rotation in RPT1 is more substantially amplified in its anticlockwise neighbouring ATPases than in its clockwise neighbours. Red arrows mark the centre of the AAA ring. **j**, Structural comparison between E_{D1} (colour) and E_{C2} (grey) in which the large AAA subdomain of RPT1 is used to align the two AAA-ring structures. A small 5° out-of-plane rotation of the large AAA subdomain of RPT1 relative to the AAA ring is observed during the re-engagement of RPT1–RPT2 with the substrate. **k**, Structural comparison between E_{D1} (colour) and E_{C2} (grey), by using the large AAA subdomain of RPT5 to align the two AAA-ring structures. A 30° out-of-plane rotation of the large AAA subdomain of RPT5 relative to the AAA ring is observed during disengagement of RPT5 from the substrate. The right panel, rotated vertically against the left panel, shows that the out-of-plane rotation in RPT5 is amplified in its anticlockwise neighbouring ATPases more substantially than in its clockwise neighbours. Red arrows mark the centre of the AAA ring. **l**, Structural comparison between E_{D2} (colour) and E_{D1} (grey) in which the large AAA subdomain of RPT5 is used to align the two AAA-ring structures. An 8° out-of-plane rotation of the large AAA subdomain of RPT5 relative to the AAA ring is observed during re-engagement of RPT5 with the substrate.

© 2019 Springer Nature Limited. All rights reserved.



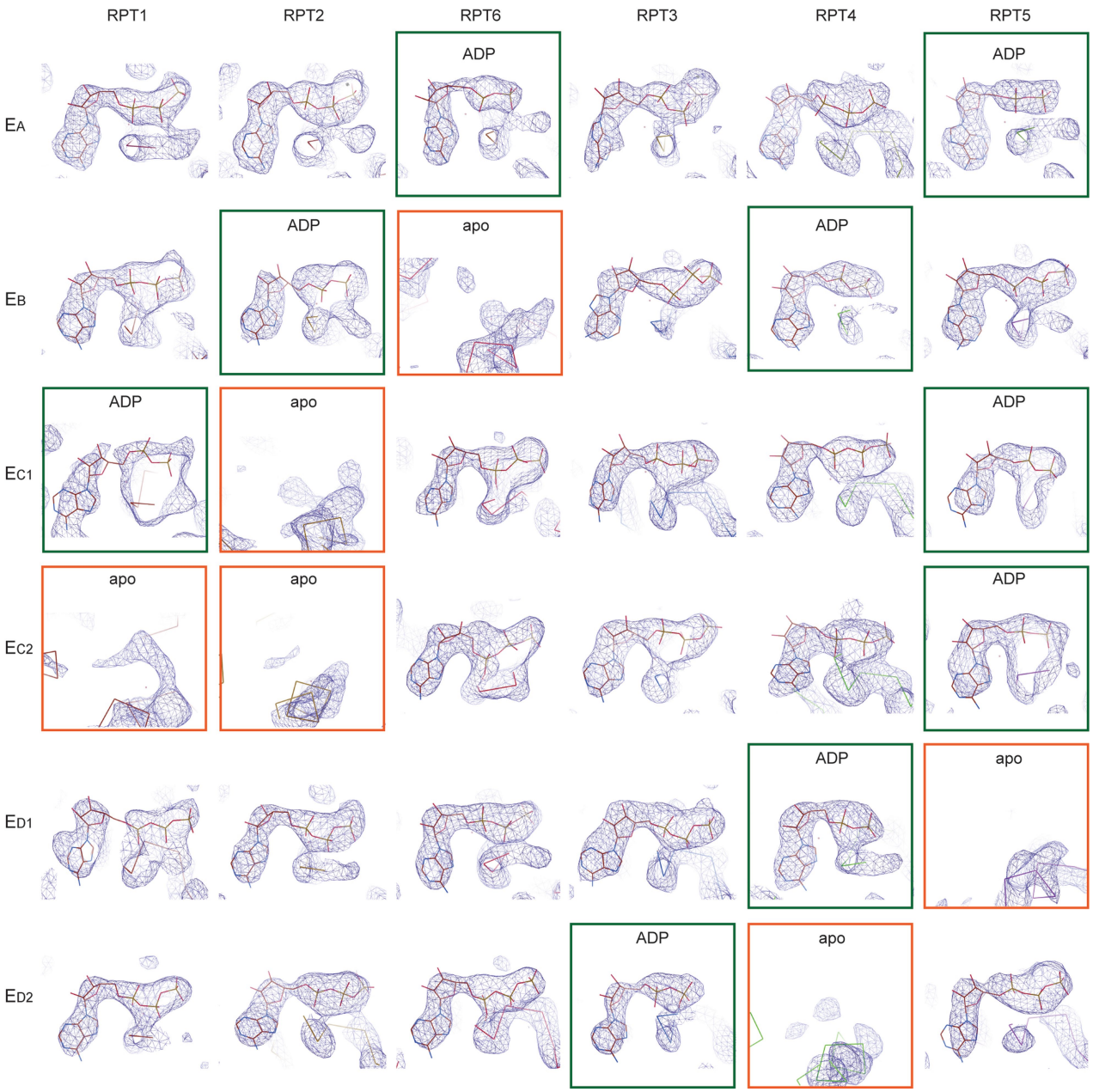
Extended Data Fig. 5 | The RP–CP interface in different states. **a**, Comparison of the RP–CP interface and RPT C-terminal tail insertions into the α-pockets of the CP in different states. The cryo-EM densities of the RP–CP interfaces are shown as a grey surface representation. The red dashed circles highlight the densities of the RPT C-terminal tails. **b**, Close-up views and comparison of the RPT C-tail densities superimposed with the atomic models in different states. The cryo-EM densities of the RPT C-tails are shown in blue mesh representation. The atomic models of the RPT C-tails are shown in stick representation. The CP structures are shown as grey cartoon representations.

© 2019 Springer Nature Limited. All rights reserved.



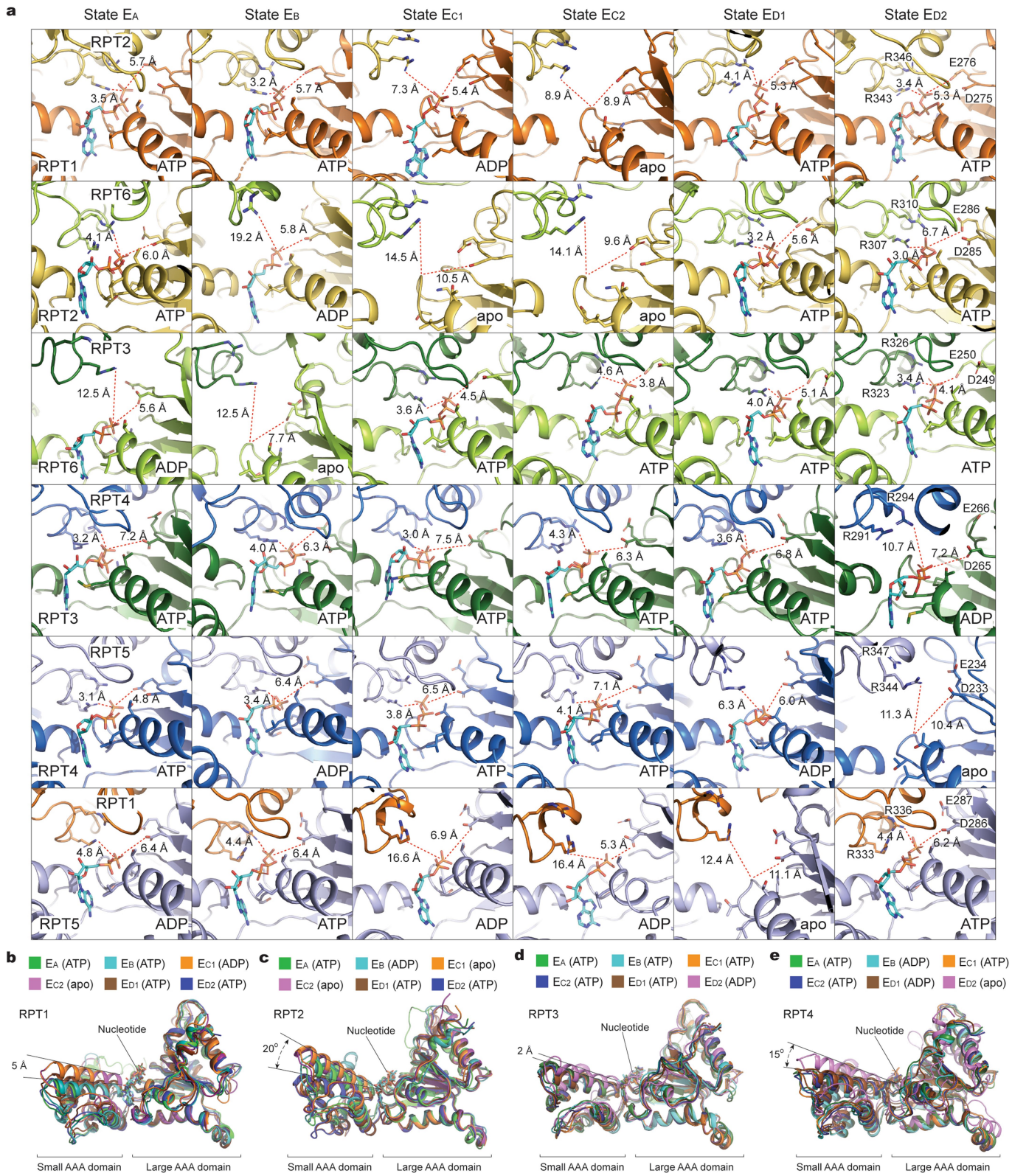
Extended Data Fig. 6 | Substrate densities in different states. a, Close-up views of two typical substrate densities observed in the CP chamber in state E_A. Left, the substrate density directly contacting the proteolytically active Thr1 in subunit β2. Right, a long substrate density at the seam between two β4 subunits inside the CP. **b**, The overall ATPase ring density of state E_B (left) and a close-up view of the substrate density (right). **c**, The overall ATPase ring density of states $E_{C1,2}$ (left) and a close-up view of the substrate density (right). **d**, The overall ATPase ring density of states $E_{D1,2}$ (left) and a close-up view of the substrate density (right). All close-up views were directly screen-copied from Coot[56] after atomic modelling into the density maps without modification.

© 2019 Springer Nature Limited. All rights reserved.



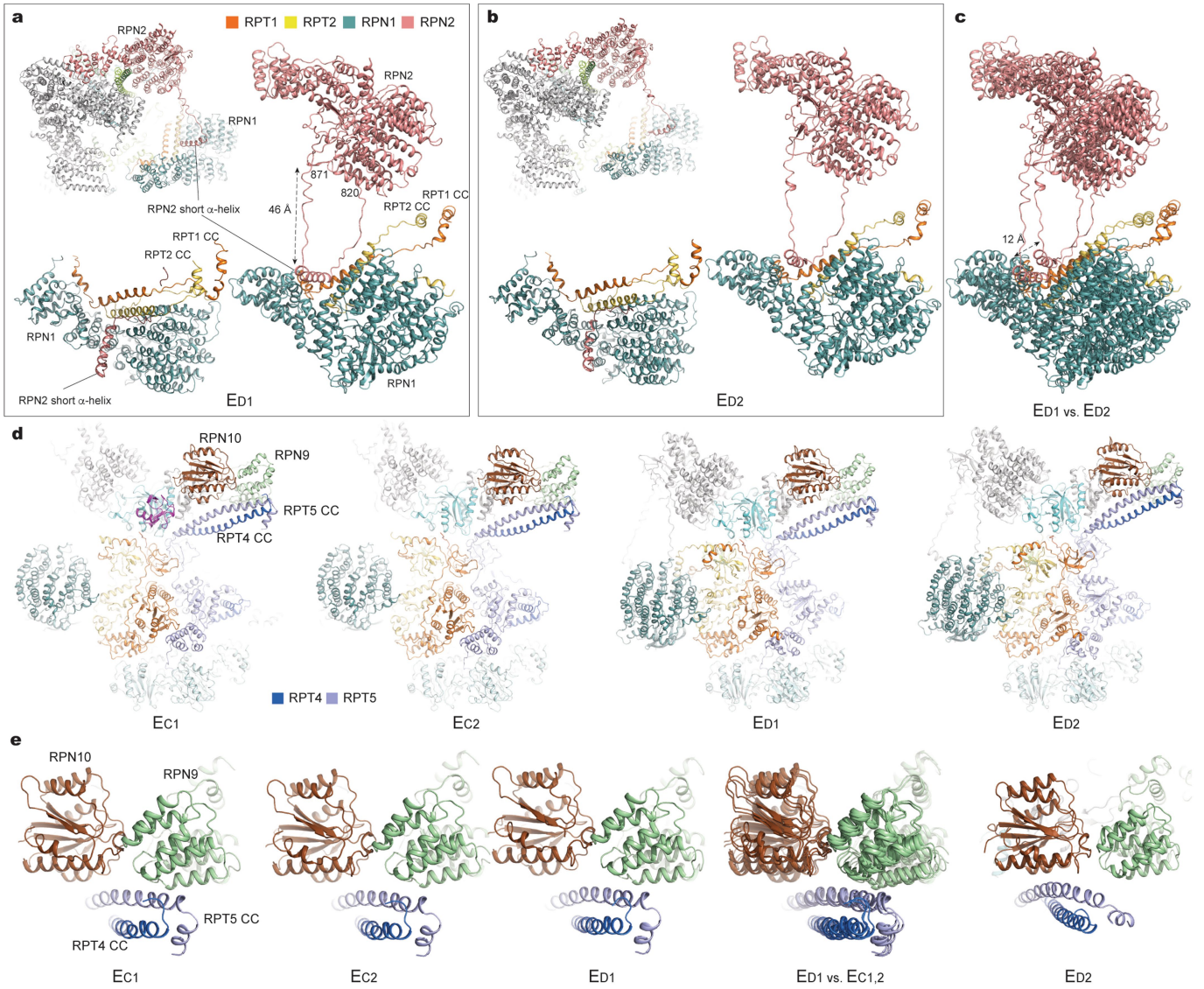
Extended Data Fig. 7 | Nucleotide densities in all states. The nucleotide densities fitting with atomic models are shown in blue mesh. All close-up views were directly screen-copied from Coot[56] after atomic modelling into the density maps without modification. At the contour level commonly used for atomic modelling, the potential nucleotide densities in the apo-like subunits mostly disappear, although they can appear as partial nucleotide shapes at a much lower contour level.

© 2019 Springer Nature Limited. All rights reserved.



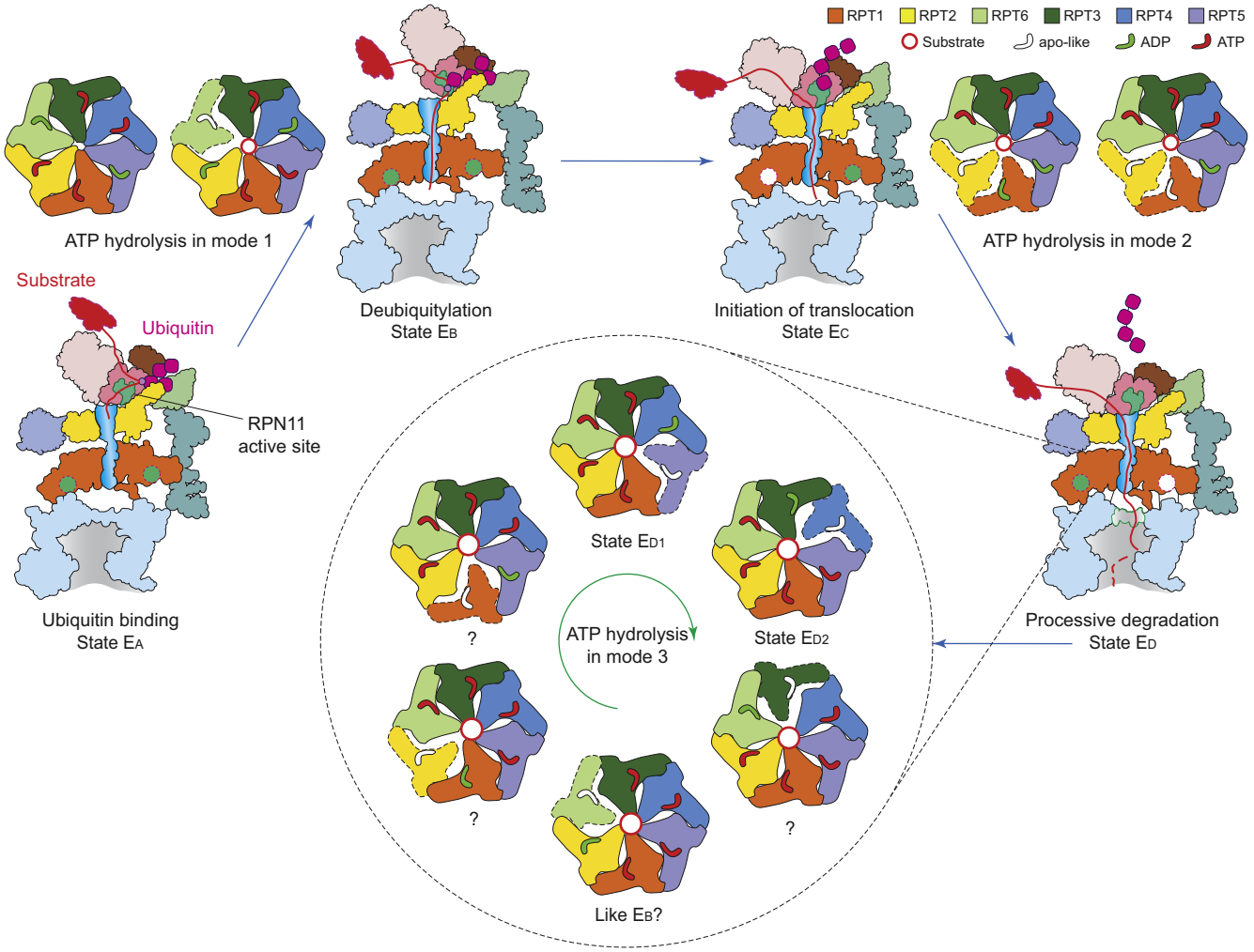
Extended Data Fig. 8 | Geometries of nucleotide-binding pockets and nucleotide-driven intrasubunit conformational changes of AAA domains. a, Comparison of the nucleotide-binding pockets of six ATPases in all states illustrates a common pattern in the geometry of the nucleotide-binding sites. Each row shows the geometry of the nucleotide-binding pocket of one ATPase in all six states. In each panel showing an ATP or ADP-bound state, one red dashed line marks the distance from the β- or γ-phosphate of the nucleotide to the arginine finger of the adjacent ATPase, and the other line marks the distance from the same phosphate to the Walker B motif. In the case of apo-like states, the red lines extend to the proline of the Walker A motif rather than to the phosphate groups. These geometries indicate the potential reactivity of these sites[33]. When the ATPase is positioned in the middle of the pore-loop staircase, but not at the lowest position, the nucleotide-binding pockets are tightly packed regardless of whether ATP or ADP is bound. By contrast, when the ATPase is either in the lowest position of the substrate-pore loop staircase or disengaged from the substrate, the nucleotide-binding pocket is rather open regardless of whether it is ADP-bound or free of nucleotide. **b**–**e**, Superpositions of the AAA domain structures of RPT1 (**b**), RPT2 (**c**), RPT3 (**d**) or RPT4 (**e**) from six distinct states aligned against their large AAA subdomains. RPT1 assumes two major conformations and RPT2 assumes three.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Changes in lid–base interactions are associated with ATP hydrolysis events through long-range allosteric regulation. **a**, **b**, Long-range association between RPN1 and RPN2 through a looping structure from RPN2 (residues 820–871) observed in states E_{D1} (**a**) and E_{D2} (**b**). **c**, Comparison of the RPN1–RPN2 long-range association between these two states shows a marked, 12 Å movement of RPN1 relative to the CP. In both states (E_{D1} and E_{D2}), the RPN1 toroidal domain and the CC domain of the RPT1–RPT2 dimer together form a surface cavity into which a short helix from RPN2 is inserted[34]. This helix resides in the middle of a long loop (residues 820–871) emanating from the toroidal domain of RPN2. The long-range association of RPN1 and RPN2 seems to stabilize a larger interface formed between RPN1–RPN2 and RPT1–RPT2. However, such a quaternary architecture is not observed in other states ($E_{A–C}$). In states $E_{C1,2}$, the RPN1 density is considerably blurred, reflecting strong motions that potentially break the long-range RPN1–RPN2 association (Fig. 1b, Extended Data Fig. 3a). Thus, the specific RPN1 conformation in each state appears to be highly coordinated with the hydrolytic cycle of the ATPase ring, and is controlled by RPN1's interactions with RPN2 in a long-range fashion. **d**, Comparison of the interactions of the CC domain of RPT4–RPT5 with RPN9 and RPN10 in states $E_{C1,2}$ and $E_{D1,2}$. **e**, Close-up views of the CC domain of RPT4–RPT5 in contact with RPN9 in states $E_{C1,2}$ and E_{D1}, and of this CC domain's contact switching to RPN10 in state E_{D2}. These observations are consistent with a recent study[35].

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Expanded model of the complete cycle of substrate processing by the human 26S proteasome. The cartoon summarizes the concept of three principal modes of coordinated ATP hydrolysis observed in the seven states and our proposal of how they regulate the complete cycle of substrate processing by the proteasome holoenzyme. Coordinated ATP hydrolysis in modes 1, 2 and 3 features hydrolytic events in two oppositely positioned ATPases[11,36], in two consecutive ATPases[9,37,38], and in only one ATPase at a time[39,40,43–46], respectively. Substrate processing undergoes three major steps before CP gate opening for processive translocation: (1) ubiquitin recognition; (2) simultaneous deubiquitylation and substrate engagement with the AAA-ATPase ring; and (3) translocation initiation, which involves multiple simultaneous events, including ubiquitin release, ATPase repositioning and switching of the RPT C-tail insertion pattern. In some cases, the initiation of translocation may precede deubiquitylation. In steps 1 and 2, the ATPases follow mode-1 ATP hydrolysis. In step 3, they follow mode-2 ATP hydrolysis. After the gate is open, the AAA-ATPases hydrolyse ATP in mode 3, in which only one nucleotide is hydrolysed at a time.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Cryo-EM data collection, refinement and validation statistics

	E_A (EMDB-9215) (PDB 6MSB)	E_{A1} (EMDB-9216,9223) (PDB 6MSB)	E_{A2} (EMDB-9217,9224) (PDB 6MSD)	E_B (EMDB-9218,9225) (PDB 6MSE)	E_{C1} (EMDB-9219,9226) (PDB 6MSG)	E_{C2} (EMDB-9220,9227) (PDB 6MSH)	E_{D1} (EMDB-9221,9228) (PDB 6MSJ)	E_{D2} (EMDB-9222,9229) (PDB 6MSK)
Data collection and processing								
Magnification	105,000	105,000	105,000	105,000	105,000	105,000	105,000	105,000
Voltage (kV)	300	300	300	300	300	300	300	300
Electron exposure (e–/Å^2)	44	44	44	44	44	44	44	44
Defocus range (μm)	-0.6 to -3.5	-0.6 to -3.5	-0.6 to -3.5	-0.6 to -3.5	-0.6 to -3.5	-0.6 to -3.5	-0.6 to -3.5	-0.6 to -3.5
Pixel size (Å)	0.685	0.685	0.685	0.685	0.685	0.685	0.685	0.685
Symmetry imposed	C1	C1	C1	C1	C1	C1	C1	C1
Initial particle images (no.)	3,584,040	3,584,040	3,584,040	3,584,040	3,584,040	3,584,040	3,584,040	3,584,040
Final particle images (no.)	184,848	105,157	79,691	242,965	112,776	71,651	288,915	348,646
Map resolution (Å)	2.8	3.0	3.2	3.3	3.5	3.6	3.3	3.2
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143
Map resolution range (Å)	2.5-5.8	2.5-5.8	2.5-6.9	2.5-4.7	2.8-5.8	3.0-6.9	2.5-5.8	2.5-4.7
Refinement								
Initial model used		5VFS						5VFO(partly)
Model resolution (Å)	3.2	3.4	3.8	3.5	3.7	3.8	3.4	3.4
FSC threshold	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Model resolution range (Å)	2.5-5.8	2.5-5.8	2.5-6.9	2.5-4.7	2.8-5.8	3.0-6.9	2.5-5.8	2.5-4.7
Map sharpening *B* factor (Å^2)	-50	-40	-46	-35	-30	-30	-35	-40
Model composition								
Non-hydrogen atoms	63654	104938	105426	105147	104415	103729	105316	105218
Protein residues	8203	13391	13418	13400	13314	13236	13420	13420
Ligands	11	12	12	9	9	5	10	11
B factors (Å^2)								
Protein	121.13	145.9	151.26	182.77	174.88	178.15	162.57	182.3
Ligand	125.87	110.13	108.00	123.85	143.96	106.39	135.00	162.17
R.m.s. deviations								
Bond lengths (Å)	0.008	0.007	0.006	0.006	0.005	0.007	0.007	0.005
Bond angles (°)	0.908	1.013	1.005	1.017	0.888	0.980	1.088	1.033
Validation								
MolProbity score	1.49	1.76	1.76	1.83	1.65	1.73	1.88	1.84
Clashscore	2.72	4.86	4.91	5.57	4.17	4.39	6.09	5.48
Poor rotamers (%)	0.47	0.54	0.42	0.46	0.4	0.62	0.73	0.69
Ramachandran plot								
Favored (%)	93.3	91.34	91.42	90.89	91.88	91.97	90.17	90.16
Allowed (%)	6.52	8.36	8.33	8.77	7.91	7.82	9.4	9.35
Disallowed (%)	0.18	0.3	0.25	0.34	0.21	0.21	0.43	0.49

Atomic model refinement with state E_A map was done only for the CP and ATPase using part of the full atomic model from state E_{A1}. The final refined atomic model, 6MSB, includes the model components of CP and ATPase refined from the combined E_A map and the rest of the holoenzyme from the E_{A1} map. 5VFO was used only for the CP component for the initial atomic model of E_{D2}.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Summary of key structural features

Key structural features connecting the seven states into a continuum of dynamic structural changes							
States	**E_{A1}**	**E_{A2}**	**E_B**	**E_{C1}**	**E_{C2}**	**E_{D1}**	**E_{D2}**
Locations of ubiquitin densities	RPN1, RPN10, RPT4/5 CC	RPN11, RPN10, RPN1, RPT4/5 CC	RPN11, RPN10, RPT4/5 CC	RPN11	None observed	None observed	None observed
RPN11 Ins-1 conformation	Large loop	β-hairpin	β-hairpin	β-hairpin	Retracted small loop	Retracted small loop	Retracted small loop
Pore-loop staircase contacting substrate	None	None	RPT3 at the top, followed by RPT4/5/1/2	RPT6 at the top, followed by RPT3/5/1	RPT6 at the top, followed by RPT3/5/1	RPT1 at the top, followed by RPT2/6/3/4	RPT5 at the top, followed by RPT1/2/6/3
RPT subunits disengaged from pore loop staircase	RPT6	RPT6	RPT6	RPT1 and RPT2	RPT1 and RPT2	RPT5	RPT4
RPT subunits with apo-like state	None	None	RPT6	RPT2	RPT1 and RPT2	RPT5	RPT4
RPT subunits with ADP bound	RPT6 and RPT5	RPT6 and RPT5	RPT2 and RPT4	RPT1 and RPT5	RPT5	RPT4	RPT3
RPT C-tails insertion into α-pockets	RPT3 and RPT5	RPT3 and RPT5	RPT2/3/5	RPT2/3/5/6	RPT2/3/5/6	RPT1/2/3/5/6	RPT1/2/3/5/6
CP gate state	Closed	Closed	Closed	Closed	Closed	Open	Open
Resemblance to the substrate-free conformations[11]	S_A	S_A (in ATPase and CP)	None	S_B (in overall lid-base relationship)	S_C (in overall lid-base relationship)	S_D (in overall lid-base relationship, RP-CP interface and CP gate)	S_{D2} (in pore-loop staircase, overall lid-base relationship, RP-CP interface and CP gate)
Functional step	Ubiquitin recognition	Ubiquitin transfer	Deubiquitylation	Initiation of translocation	Priming for CP gate opening	Processive translocation	Processive translocation

Potential substrate-binding sites observed in the CP chamber			
State	**Key contacts at the binding site**	**Features**	**Number of substrate residues**
E_A, E_B, $E_{C1,2}$, $E_{D1,2}$	Thr1, Cys31 of β2, Cys129 of β3	Symmetric	6
E_A	Asn24, Tyr134, Phe137 of β4	At the seam between two β4 subunits	10
E_A	Tyr103 of α1, Tyr61, Tyr90 of β1, and Phe88 of β2	Symmetric, at the inter-subunit interface	5
E_A	Asn90 of α5, Tyr90 of β5, Phe101 of β6	Symmetric, at the inter-subunit interface	3
E_A, E_B, $E_{C1,2}$, $E_{D1,2}$	Tyr105, Arg117 of α1, His88 of α2	Symmetric	3
E_A, E_B, $E_{C1,2}$, $E_{D1,2}$	Tyr59, Cys91, Tyr98 of β4	Symmetric	3
E_A	Phe69, Cys91 and Tyr98 of β3	Asymmetric, only present in the chamber with the CP gate open	3
E_A	Ile3, Tyr6, Tyr104 of β3, Tyr120 of β4	Asymmetric, only present in the chamber with the CP gate open	3
E_B, $E_{C1,2}$, $E_{D1,2}$	Tyr30 of β7	Symmetric	4

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0778-7

Direct observation of incommensurate magnetism in Hubbard chains

Guillaume Salomon[1]*, Joannis Koepsell[1], Jayadev Vijayan[1], Timon A. Hilker[1], Jacopo Nespolo[2,3], Lode Pollet[2], Immanuel Bloch[1,2] & Christian Gross[2]

The interplay between magnetism and doping is at the origin of exotic strongly correlated electronic phases and can lead to novel forms of magnetic ordering. One example is the emergence of incommensurate spin-density waves, which have wavevectors that do not belong to the reciprocal lattice. In one dimension this effect is a hallmark of Luttinger liquid theory, which also describes the low-energy physics of the Hubbard model[1]. Here we use a quantum simulator that uses ultracold fermions in an optical lattice[2–8] to directly observe such incommensurate spin correlations in doped and spin-imbalanced Hubbard chains using fully spin- and density-resolved quantum gas microscopy. Doping is found to induce a linear change in the spin-density wavevector, in excellent agreement with predictions from Luttinger theory. For non-zero polarization we observe a reduction in the wavevector with magnetization, as expected from the antiferromagnetic Heisenberg model in a magnetic field. We trace the microscopic-scale origin of these incommensurate correlations to holes, doublons (double occupancies) and excess spins, which act as delocalized domain walls for the antiferromagnetic order. In addition, by inducing interchain coupling we observe fundamentally different spin correlations around doublons and suppression of incommensurate magnetism at finite (low) temperature in the two-dimensional regime[9]. Our results demonstrate how access to the full counting statistics of all local degrees of freedom can be used to study fundamental phenomena in strongly correlated many-body physics.

One-dimensional (1D) quantum systems are paradigmatic examples of the breakdown of Landau Fermi-liquid theory. The free quasiparticles that are present in higher dimensions are replaced by collective excitations, leading to striking phenomena such as spin–charge separation[1]. Luttinger liquid theory[10] generically describes the low-energy physics of gapless 1D systems ranging from quasi-1D conductors and spin liquids to chiral edge modes in the fractional quantum Hall effect[11]. In particular, the repulsive single-band Hubbard model, which provides a minimal microscopic-scale description of doped antiferromagnets, can be described through this approach. Away from half-filling, Luttinger liquid theory predicts incommensurate magnetism with an algebraically decaying incommensurate spin-density wave (SDW) at zero temperature, whose vector varies linearly with density[1]. Also, the presence of a spin imbalance in the 1D Hubbard model can lead to incommensurate spin correlations[12]. Short-range incommensurate magnetism is expected to survive at finite temperature, where conformal field theory predicts an exponential decay of the spin correlations with distance[13]. Luttinger liquids have been experimentally studied in traditional condensed matter systems such as carbon nanotubes via conductance and scanning tunnelling microscopy measurements[14,15]. In particular, magnetism on weakly coupled quasi-1D spin-1/2 chains[16,17] and on ladder systems[18] has been studied using neutron scattering. In higher dimensions, incommensurate SDWs have been detected in underdoped regions of certain high-temperature superconductors via neutron scattering[19]. An interpretation of the results in terms of holes organized in stripes has been proposed, which results in an effective 1D description of the higher-dimensional systems in which the stripes form domain walls in the antiferromagnet. Here we use real-space spin- and density-resolved quantum gas microscopy to directly study the effects of both doping and polarization on finite-range spin correlations in the 1D Hubbard model. We measure the linear change in the SDW vector as a function of density, in excellent agreement with quantum Monte Carlo (QMC) calculations. In the presence of a spin population imbalance, we observe an increase of the SDW wavelength with polarization, as predicted by Luttinger liquid theory and in good agreement with exact diagonalization calculations of the Heisenberg chain. Finally, we report on the evolution of antiferromagnetic spin correlations around doublons in the crossover from one to two dimensions. We find the magnetic environment around doublons to change fundamentally when spin correlations appear in the transverse direction, suggesting the formation of a magnetic polaron[9].

Our experiments started by loading a balanced two-dimensional (2D) degenerate spin mixture of ^{6}Li atoms in the two lowest Zeeman

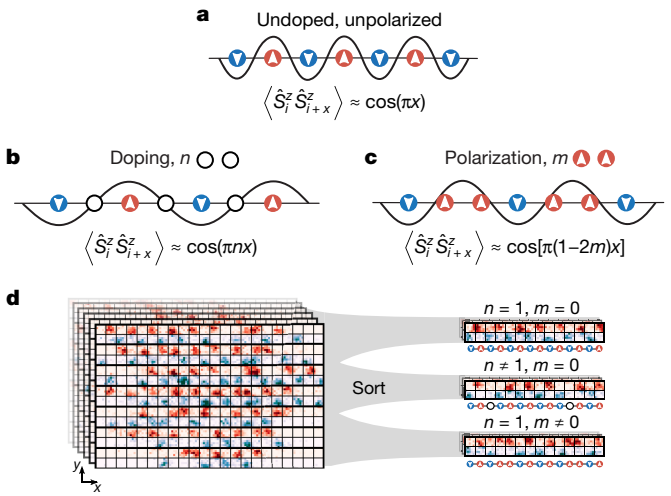


Fig. 1 | Probing incommensurate spin correlations in Hubbard chains. **a**, Spin correlations in spin-balanced Hubbard chains at half filling ($n = 1$) form at a commensurate wavevector π. Red (blue) circles with white up (down) arrows denote up (down) spins. **b**, When the system is doped ($n \neq 1$), incommensurate spin correlations at wavevector πn develop owing to delocalized holes and doublons, which increase the distance between antiferromagnetically correlated spins. **c**, At finite polarization $m \neq 0$, incommensurate spin correlations at a wavevector $\pi(1 - 2m)$ arise owing to excess spins. **d**, Left, single-spin and density-resolved experimental images, each containing seven independent Hubbard chains along y, separated by thick lines, where up (down) spins are represented in red (blue). Right, in post-analysis, we group the data by polarization and doping to analyse their individual effect on spin correlations along x.

[1]Max-Planck-Institut für Quantenoptik, Garching, Germany. [2]Fakultät für Physik, Ludwig-Maximilians-Universität, Munich, Germany. [3]INO-CNR BEC Center and Dipartimento di Fisica, Universita di Trento, Povo, Italy. *e-mail: guillaume.salomon@mpq.mpg.de

© 2019 Springer Nature Limited. All rights reserved.

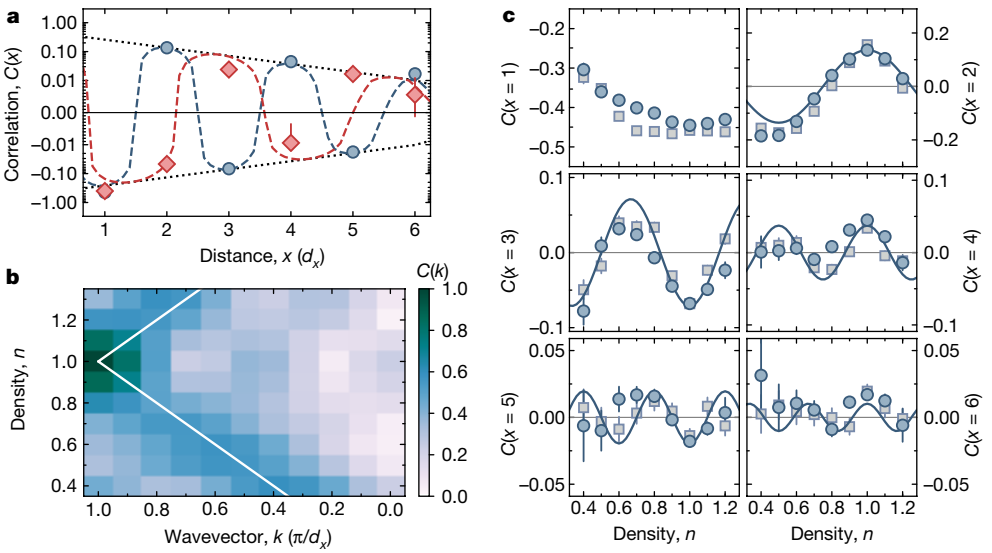


Fig. 2 | Incommensurate spin correlations versus doping. a, Spin correlations $C(x)$ at half-filling (blue) and at $n = 0.7$ (red). The dotted lines show the decay obtained from an exponential fit of the rectified spin correlations $(-1)^x C(x)$ at half-filling. The dashed lines are the Luttinger liquid theory predictions, calculated using the amplitude and decay length obtained from a fit of the $n = 1$ experimental data. The sign change observed for $d \geq 2$ in the doped case originates from delocalized holes stretching the distance between antiferromagnetically correlated spins. **b**, Away from half-filling, the normalized Fourier transform of the spin correlations $C(k)$ reveals a linear increase of the SDW wavevector with density. The white line is the Luttinger liquid theory result $k_{\mathrm{SDW}} = \pi n$. **c**, Spin correlations $C(x)$ versus density at fixed distances $x = 1, \ldots, 6$ (blue dots) compared to QMC calculations at $T = 0.29t_x$ (grey squares). The measured densities are binned in intervals of 0.1. The blue lines are the Luttinger liquid theory prediction with wavevectors πn, calculated using the amplitude and decay length extracted from the fit in **a**. Error bars in all panels denote one standard error of the mean.

states, $|\uparrow\rangle$ and $|\downarrow\rangle$, into an optical lattice formed by two standing waves with period $d_x = 1.15\ \mu$m in the x direction and $d_y = 2.3\ \mu$m in the y direction (Fig. 1)[5]. The atoms were trapped in a single plane of a vertical lattice with 3.1 μm spacing and a depth of $17E_r^z$ where E_r^i denotes the recoil energy in direction i. The nearest-neighbour tunnelling rates were set to $t_x/h = 410$ Hz at a lattice depth of $5E_r^x$ and $t_y/h = 1.2$ Hz at $27E_r^y$ to study the 1D Hubbard model (h is the Planck constant). By decreasing the lattice depth in the y direction and ramping up the x lattice power to vary t_y/t_x, we can explore the Hubbard model through the dimensional crossover from one to two dimensions. The on-site interaction U was controlled using the broad Feshbach resonance located at 834.1 G and set to $U = 7t_x$ in the 1D regime. We directly measured the occupation and spin on each lattice site by first freezing the atomic motion before a local Stern–Gerlach-like splitting of the spin components in a superlattice along y (see ref. [5] and Fig. 1). Finally we detected the atoms via Raman sideband cooling[20]. Thanks to the ultimate resolution of our detection technique, which can detect single atoms and spins, we are able to group our data according to the total spin $S^z = (N_\uparrow - N_\downarrow)/2$ and total atom number $N = N_\uparrow + N_\downarrow$, that is, the sum of the numbers of atoms with up and down spins in each chain. These conserved quantities fluctuate for different chains and experimental runs (see Extended Data Fig. 1); however, data grouping allows us to explore the effect of doping and spin imbalance individually (see Fig. 1).

We first study the evolution of antiferromagnetic spin correlations along 1D chains as a function of doping. The correlations are quantified by the two-point correlation function

$$C(x) = 4\langle S_i^z S_{i+x}^z \rangle_{\bullet_i \bullet_{i+x}}$$

conditioned on sites i and $i + x$ being singly occupied (filled circles). Experimentally, we prepared Hubbard chains with up to $N = 23$ atoms and post-selected the experimental outcomes to the $S^z = 0$ sector to first consider the effects of doping only. Owing to the underlying harmonic confinement, the atomic cloud is inhomogeneous; using a local density approximation we define the density n as the mean occupation calculated over the sites connecting i to $i + x$ for each value of N (see Methods). From Luttinger liquid theory one expects the wavevector of the SDW to be $k_{\mathrm{SDW}} = 2k_{\mathrm{F}} = \pi n$, where k_{F} is the Fermi wavevector. At finite temperature and large distances $x \gtrsim k_{\mathrm{F}}^{-1}$, the spin correlations are predicted to decay exponentially[1]:

$$C(x) \approx A\mathrm{e}^{-x/\xi}\cos(\pi n x) \quad (1)$$

where A is a non-universal constant and ξ is the temperature-dependent correlation length that varies[1] weakly with density at $U/t = 7$. We determined A and ξ from an exponential fit of $C(x)$ at half-filling ($n = 1$) for $x = 2, \ldots, 6$, which yields $A = 0.49(4)$ and $\xi = 1.6(1)$ (Fig. 2a), where all distances are expressed in units of the lattice constant d_x (uncertainties denote one standard error of the mean). Away from half-filling, we observe a linear increase of the SDW vector for both hole and charge doping, as revealed by a Fourier transform of the rescaled spin correlation $C(k) = F\{A^{-1}\mathrm{e}^{x/\xi}C(x)\}$ (Fig. 2b). For a quantitative comparison with theory, we show in Fig. 2c the spin correlations $C(x)$ as a function of density n together with QMC calculations for a homogeneous system at temperature $T = 0.29t_x$, as well as the long-distance Luttinger prediction of equation (1). The spin correlations are found to oscillate with a periodicity of $k_{\mathrm{SDW}} = \pi n$, as expected from Luttinger theory. We attribute the microscopic-scale origin of the incommensurate correlations to delocalized doublons and holes, which increase the distance between antiferromagnetically correlated spins[21,22] and thus the wavelength of the SDW. The remaining discrepancies between the experiment, QMC calculations and the Luttinger liquid theory expectations, which are mostly visible at large distances and low densities, are attributed to the trap that induces inhomogeneous density profiles and to averaging over different chain lengths, leading to corrections to the exponential decay.

Incommensurate spin correlations are also expected to appear in the 1D Hubbard model when a spin imbalance is introduced. To isolate the effect of polarization from the influence of doping, we consider the connected two-point spin correlations $C(\tilde{x}) = 4(\langle S_{\tilde{i}}^z S_{\tilde{i}+\tilde{x}}^z \rangle - \langle S_{\tilde{i}}^z \rangle \langle S_{\tilde{i}+\tilde{x}}^z \rangle)$ in squeezed space (the tilde notation refers to quantities in squeezed space), which are obtained by removing holes and doublons from the chain in the post-analysis[22]. In squeezed space[23,24] and for large U/t_x, the system is described by a spin-1/2 antiferromagnetic Heisenberg

© 2019 Springer Nature Limited. All rights reserved.

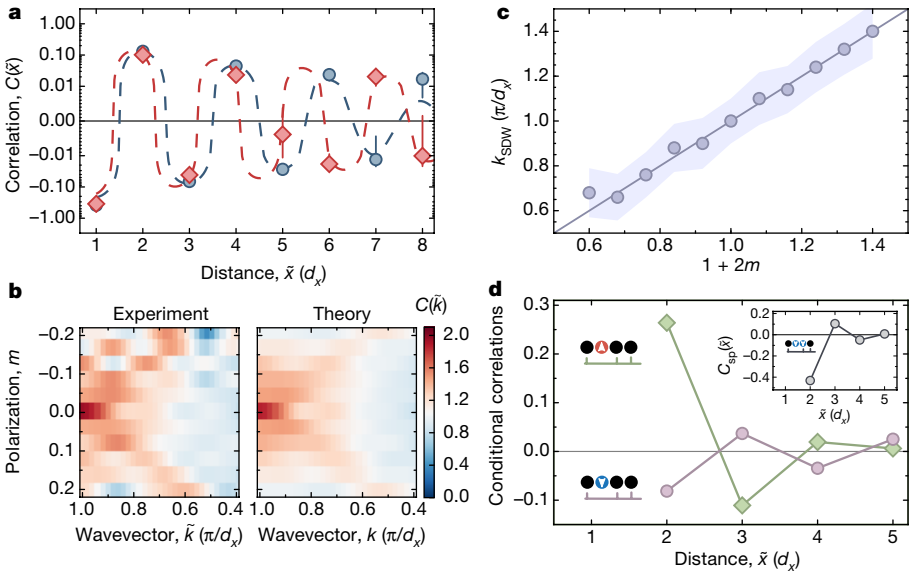


Fig. 3 | Incommensurate spin correlations versus polarization. a, Spin correlations in squeezed space $C(\tilde{x})$ for $m = 0$ (blue) and $m = 0.08$ (red). A sign change is visible at distance $\tilde{x} > 5$, reflecting an increase of the SDW wavelength away from $m = 0$. Dashed lines are fits to the Luttinger liquid theory expectation (equation (2)). **b**, Left, normalized Fourier transform, $C(\tilde{k})$, of $C(\tilde{x})$. The plot exhibits two branches departing from $k = \pi$ when $m \neq 0$, in good agreement with exact diagonalization calculations of the Heisenberg chain at $T = 0.7J$ averaged over the experimental $\{S^z, N_s\}$ (right). The binning of the polarization is in intervals of 0.04. **c**, A linear fit of the branches in **b** yields $k_{SDW} = 1.0(1) \times (1 + 2m)\pi$, in excellent agreement with the Luttinger liquid theory prediction. The shaded region denotes the Fourier-limited systematic error. **d**, Conditional spin correlations across majority $C_{maj}(\tilde{x})$ (violet circles) and minority $C_{min}(\tilde{x})$ (green diamonds) spins for $m = -0.12$. A phase shift of π is visible for the spin correlations across majority spins. Inset, conditional spin correlations across pairs of parallel spins, $C_{sp}(\tilde{x})$, revealing their domain-wall nature in a squeezed space that lies at the origin of the SDW wavelength extension. Error bars denote one standard error of the mean; in **d**, they are smaller than the symbols.

model at a polarization $m = S^z/N_s$, where N_s is the number of singly occupied sites (singlons; see Methods). For the Heisenberg chain, Luttinger liquid theory predicts incommensurate spin correlations that are linear with polarization[1] m at large distances:

$$C(\tilde{x}) \approx A_m e^{-\tilde{x}/\xi_m} \cos[\pi(1+2m)\tilde{x}] \quad (2)$$

where A_m and ξ_m are the polarization- and temperature-dependent amplitude and correlation length, respectively. The SDW wavelength measured by $C(\tilde{x})$ is thus expected to increase away from $m = 0$ because the fixed sampling at the lattice period makes $m > 0$ and $m < 0$ symmetric. Finite-size effects and the dependence[25] of ξ_m on m influence the functional form of the decay (see Methods); therefore we concentrate here on the SDW vector variation predicted by equation (2). In Fig. 3a we show $C(\tilde{x})$ for two polarizations of the chain, $m = 0$ and $m = 0.08$. We observe sign changes in $C(\tilde{x})$ for $\tilde{x} > 5$, which indicate a wavelength extension of the SDW for $m = 0.08$ compared to $m = 0$. To detect the SDW vector variation as a function of polarization without prior assumption of any functional form, we compute the zero-padded Fourier transform of the rescaled spin correlations in squeezed space, $C(\tilde{k}) = F\{C(\tilde{x})/|C(1)|\}$, for each m. It reveals two branches away from $k = \pi$ (see Fig. 3b), for $m > 0$ and $m < 0$, in qualitative agreement with the results of the exact diagonalization of the Heisenberg chain at $T = 0.7J$, where J is the exchange coupling, averaged over our experimental distribution $\{S^z, N_s\}$. In Fig. 3c a linear fit of these branches yields $k_{SDW} = 1.0(1) \times (1 + 2m)\pi$, in remarkable agreement with the Luttinger liquid theory prediction (equation (2)). Similarly to the doped case[21], we now study the microscopic-scale origin of these incommensurate spin correlations. We analyse the spin environment around the majority and minority spins

$$C_{maj}(\tilde{x}) = 4\langle S^z_{\tilde{\imath}} S^z_{\tilde{\imath}+\tilde{x}} \rangle_{S^z\sigma_{\tilde{\imath}+1}>0}$$
$$C_{min}(\tilde{x}) = 4\langle S^z_{\tilde{\imath}} S^z_{\tilde{\imath}+\tilde{x}} \rangle_{S^z\sigma_{\tilde{\imath}+1}<0}$$

by measuring the conditional expectation value of the spin correlations in squeezed space for distances $\tilde{x} \geq 2$. The conditioning $S^z\sigma_{\tilde{\imath}+1} > 0$ ($S^z\sigma_{\tilde{\imath}+1} < 0$) means that the spin $\sigma_{\tilde{\imath}+1}$ on site $\tilde{\imath}+1$ is parallel (antiparallel) to the chain magnetization S^z (Fig. 3d). The SDW measured by the correlator $C(\tilde{x})$ is the polarization-dependent weighted average of these two components (see Methods). In Fig. 3d we observe a phase shift of π in the short-distance oscillating part of $C_{maj}(\tilde{x})$, whereas $C_{min}(\tilde{x})$ stays in phase with the unpolarized case. This implies that the excess spins mostly form pairs of parallel majority spins, shifting the antiferromagnetic correlations by one lattice site in the $m \neq 0$ case. In the inset of Fig. 3d we explicitly evaluate the spin correlations across pairs of parallel spins, finding strong antiferromagnetic correlations with opposite parity compared to the unpolarized situation. This demonstrates that excess spins act as domain walls in squeezed space. Similar to the holes in the doped case, the main effect of excess spins is therefore to increase the distance between antiferromagnetically correlated spins, resulting in an increase of the SDW wavelength, as measured by $C(\tilde{x})$. Polarized synthetic Hubbard models have recently been studied also in two dimensions and the emergence of anisotropic spin correlations has been observed[8], but with an unchanged wavevector.

We now explore the evolution of the spin correlations in the 1D–2D crossover, a situation relevant to quasi-1D antiferromagnets[17]. Whereas in 1D there is no magnetic energy cost associated with the delocalization of holes and doublons, this phenomenon is expected to break down in higher dimensions. In a 2D antiferromagnetic background the motion of holes and doublons leads to strings of flipped spins, resulting in the confinement of spin and charge[9,26]. The antiferromagnetic spin correlations around doublons and holes, which are at the origin of the SDW wavelength extension in the doped 1D regime, are therefore expected to qualitatively change in the crossover from one to two dimensions[9]. We prepared 2D clouds with up to 70 atoms and studied spin correlations while varying t_y/t_x between 0 and 1 and keeping $U/t_x = 14$ constant (see Methods). When increasing t_y/t_x, we first observe a decrease in the amplitude of the spin correlations

$$C(x, y) = 4\langle S^z_{i,j} S^z_{i+x,j+y} \rangle_{\bullet_{i,j}\bullet_{i+x,j+y}}$$

© 2019 Springer Nature Limited. All rights reserved.

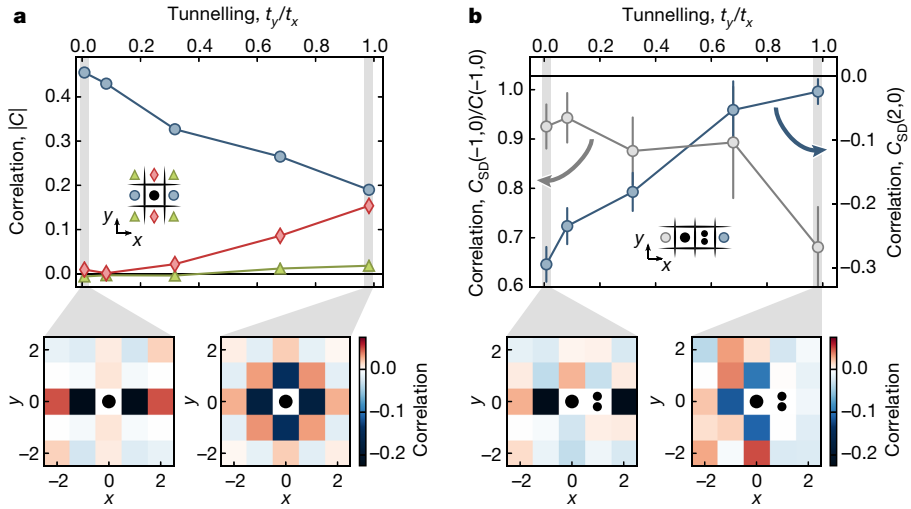


Fig. 4 | Spin correlations in the 1D–2D crossover. a, Spin correlations $C(x, y)$ as a function of the ratio t_y/t_x: $|C(1, 0)|$ (blue circles), $|C(0, 1)|$ (red diamonds) and $|C(1, 1)|$ (green triangles) at $U/t_x = 14$. The spin correlations along x decrease as spin correlations develop in the y direction. The lower panels show the 2D spin correlation amplitudes $C(x, y)$ in the 1D (left) and 2D (right) limits. **b**, Spin correlations across doublons $C_{SD}(2, 0)$ (blue) and next to doublons $C_{SD}(-1, 0)/C(-1, 0)$ (grey) along the x direction. Antiferromagnetic correlations across doublons, at the origin of incommensurate correlations in the 1D limit, are strongly suppressed as spin correlations develop in two dimensions. In this limit the delocalization of doublons also leads to a reduction of antiferromagnetic correlations on neighbouring sites, suggesting the formation of a magnetic polaron. The lower panels show the spin correlations $C_{SD}(x, y)$ between sites (0, 0) (black circle) and (x, y) conditioned on finding a doublon on site (1, 0) (double black circle) in the 1D (left) and 2D (right) case. Error bars in all panels denote one standard error of the mean.

along x and the emergence of spin correlations in the transverse directions (Fig. 4a)[27]. This decrease is expected even at zero temperature and half-filling, where the nearest-neighbour spin correlations $C(1)$ change from -0.6 to -0.36 owing to the higher coordination number modifying the quantum fluctuations[4].

Next, we study the magnetic environment around doublons in the dimensional crossover from one to two dimensions through

$$C_{SD}(x, y) = 4\langle S^z_{i,j} S^z_{i+x,j+y}\rangle_{\bullet_{i,j}\circ_{i+1,j}\bullet_{i+x,j+y}}$$

where the open circle denotes a doublon located at site $(i + 1, j)$ (see Methods). We find that the spin correlations across doublons, $C_{SD}(2, 0)$, are strongly suppressed while 2D spin correlations develop, which is in stark contrast to the 1D case (Fig. 4b). Owing to the harmonic confinement, the few double occupancies are located at the centre of the trap, where the average density is highest and where magnetic correlations are expected to compete with doublon delocalization. In addition to the vanishingly small antiferromagnetic correlations across doublons, we observe a reduction in the nearest-neighbour spin correlations in their vicinity, $C_{SD}(-1, 0)/C(-1, 0)$, to about 70% of the undoped case (Fig. 4b). This suggests the formation of a magnetic polaron[9], which in the extreme limit $U/t \to \infty$ corresponds to the Nagaoka polaron[28]. These local observations provide an initial microscopic-scale picture of the underlying process and explain the recent observation of the global suppression of long-range antiferromagnetic order in two dimensions away from half-filling[29].

Through the direct simultaneous measurement of both density and spin in the doped and spin-imbalanced 1D Hubbard model, we shed light on the connection between incommensurate spin correlations and microscopic-scale degrees of freedom. The spin environment around doublons was found to differ drastically in the 1D and 2D cases, calling for further experimental studies of the formation of magnetic polarons in homogeneous systems[9,30]. Possible routes towards the unambiguous identification of a polaron include the detection of its size using spin correlations and a measurement of its effective mass dynamically or spectroscopically. Another interesting extension of this work is the study of spin correlations as a function of the number of coupled chains, where the parity of the latter is predicted to lead to striking differences between even and odd cases, similarly to the problem of half-integer and integer spin chains[31,32]. At low enough temperature the study of spin and density correlations in hole-doped coupled chains is also expected to reveal a binding of holes to form stripes, which directly extends the domain-wall concept discussed here to two dimensions[33]. A study of such effects through quantum gas microscopy can provide microscopic-scale insights into the physics of the doped repulsive Hubbard model.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0778-7.

Received: 23 March 2018; Accepted: 12 October 2018;
Published online 12 December 2018.

1. Giamarchi, T. *Quantum Physics in One Dimension* (Clarendon Press, Oxford, 2003).
2. Greif, D., Uehlinger, T., Jotzu, G., Tarruell, L. & Esslinger, T. Short-range quantum magnetism of ultracold fermions in an optical lattice. *Science* **340**, 1307–1310 (2013).
3. Hart, R. A. et al. Observation of antiferromagnetic correlations in the Hubbard model with ultracold atoms. *Nature* **519**, 211–214 (2015).
4. Parsons, M. F. et al. Site-resolved measurement of the spin-correlation function in the Fermi–Hubbard model. *Science* **353**, 1253–1256 (2016).
5. Boll, M. et al. Spin- and density-resolved microscopy of antiferromagnetic correlations in Fermi–Hubbard chains. *Science* **353**, 1257–1260 (2016).
6. Cheuk, L. W. et al. Observation of spatial charge and spin correlations in the 2D Fermi–Hubbard model. *Science* **353**, 1260–1264 (2016).
7. Drewes, J. H. et al. Antiferromagnetic correlations in two-dimensional fermionic Mott-insulating and metallic phases. *Phys. Rev. Lett.* **118**, 170401 (2017).
8. Brown, P. T. et al. Spin-imbalance in a 2D Fermi–Hubbard system. *Science* **357**, 1385–1388 (2017).
9. Dagotto, E. Correlated electrons in high-temperature superconductors. *Rev. Mod. Phys.* **66**, 763–840 (1994).
10. Haldane, F. D. M. 'Luttinger liquid theory' of one-dimensional quantum fluids. I. Properties of the Luttinger model and their extension to the general 1D interacting spinless Fermi gas. *J. Phys. C Solid State Phys.* **14**, 2585–2609 (1981).
11. Wen, X. G. Chiral Luttinger liquid and the edge excitations in the fractional quantum Hall states. *Phys. Rev. B* **41**, 12838–12844 (1990).
12. Frahm, H. & Korepin, V. E. Correlation functions of the one-dimensional Hubbard model in a magnetic field. *Phys. Rev. B* **43**, 5653–5662 (1991).
13. Cardy, J. *Scaling and Renormalization in Statistical Physics* (Cambridge Univ. Press, Cambridge, 1996).
14. Bockrath, M. et al. Luttinger-liquid behaviour in carbon nanotubes. *Nature* **397**, 598–601 (1999).
15. Lee, J. et al. Real space imaging of one-dimensional standing waves: direct evidence for a Luttinger liquid. *Phys. Rev. Lett.* **93**, 166403 (2004).

© 2019 Springer Nature Limited. All rights reserved.

16. Stone, M. B. et al. Extended quantum critical phase in a magnetized spin-1/2 antiferromagnetic chain. *Phys. Rev. Lett.* **91**, 037205 (2003).
17. Lake, B., Tennant, D. A., Frost, C. D. & Nagler, S. E. Quantum criticality and universal scaling of a quantum antiferromagnet. *Nat. Mater.* **4**, 329–334 (2005).
18. Klanjšek, M. et al. Controlling Luttinger liquid physics in spin ladders under a magnetic field. *Phys. Rev. Lett.* **101**, 137207 (2008).
19. Tranquada, J. M., Sternlieb, B. J., Axe, J. D., Nakamura, Y. & Uchida, S. Evidence for stripe correlations of spins and holes in copper oxide superconductors. *Nature* **375**, 561–563 (1995).
20. Omran, A. et al. Microscopic observation of Pauli blocking in degenerate fermionic lattice gases. *Phys. Rev. Lett.* **115**, 263001 (2015).
21. Hilker, T. A. et al. Revealing hidden antiferromagnetic correlations in doped Hubbard chains via string correlators. *Science* **357**, 484–487 (2017).
22. Kruis, H. V., McCulloch, I. P., Nussinov, Z. & Zaanen, J. Geometry and the hidden order of Luttinger liquids: the universality of squeezed space. *Phys. Rev. B* **70**, 075109 (2004).
23. Ogata, M. & Shiba, H. Bethe–Ansatz wave function, momentum distribution, and spin correlation in the one-dimensional strongly correlated Hubbard model. *Phys. Rev. B* **41**, 2326–2338 (1990).
24. Woynarovich, F. Excitations with complex wavenumbers in a Hubbard chain. I. States with one pair of complex wavenumbers. *J. Phys. C* **15**, 85–96 (1982).
25. Bogoliubov, N. M., Izergin, A. G. & Korepin, V. E. Critical exponents for integrable models. *Nucl. Phys. B* **275**, 687–705 (1986).
26. Brinkman, W. F. & Rice, T. M. Single-particle excitations in magnetic insulators. *Phys. Rev. B* **2**, 1324–1338 (1970).
27. Greif, D., Jotzu, G., Messer, M., Desbuquois, R. & Esslinger, T. Formation and dynamics of antiferromagnetic correlations in tunable optical lattices. *Phys. Rev. Lett.* **115**, 260401 (2015).
28. White, S. R. & Affleck, I. Density matrix renormalization group analysis of the Nagaoka polaron in the two-dimensional *t*–*J* model. *Phys. Rev. B* **64**, 024411 (2001).
29. Mazurenko, A. et al. A cold-atom Fermi–Hubbard antiferromagnet. *Nature* **545**, 462–466 (2017).
30. Grusdt, F. et al. Parton theory of magnetic polarons: mesonic resonances and signatures in dynamics. *Phys. Rev. X* **8**, 011046 (2017).
31. Dagotto, E. & Rice, T. M. Surprises on the way from one- to two-dimensional quantum magnets: the ladder materials. *Science* **271**, 618–623 (1996).
32. White, S. R. & Scalapino, D. J. Hole and pair structures in the *t*–*J* model. *Phys. Rev. B* **55**, 6504–6517 (1997).
33. White, S. R. & Scalapino, D. J. Density matrix renormalization group study of the striped phase in the 2D *t*–*J* model. *Phys. Rev. Lett.* **80**, 1272–1275 (1998).

Acknowledgements We thank T. Giamarchi for exchanges on incommensurate magnetism, D. Huse, A. Recati, E. Demler and F. Grusdt for discussions and P. Sompet for reading the manuscript. Financial support was provided by the Max Planck Society (MPG) and the European Union (UQUAM, QSIMGAS, MIR-BOSE), and J.K. acknowledges funding from the Hector Fellow Academy.

Reviewer information *Nature* thanks M. Lewenstein, C. de Morais Smith and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions G.S., T.A.H., J.K., J.V., I.B. and C.G. planned the experiment and analysed and discussed the data. J.N. and L.P. performed the QMC simulations. All authors contributed to the interpretation of the data and the writing of the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0778-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to G.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

Ultracold lattice gas preparation. The experimental protocol used in the experiments reported here closely followed our previous work[21]. Our experiments started with a degenerate spin mixture of ^{6}Li atoms in the lowest two Zeeman states, $|\pm\rangle = |F = 1/2, m_F = \pm 1/2\rangle$ (where F and m_F define the hyperfine state) trapped in a single plane of a vertical optical lattice. The lattice spacing was 3.1 μm and the depth was $17E_r^z$ ($27E_r^z$) in the 1D (crossover) case, where $E_r^i = h^2/(8md_i^2)$ is the recoil energy, m the atomic mass and d_i the lattice spacing along direction i. The total atom number N of the cloud was tuned by varying the depth of a radial trap at the endpoint of the evaporative cooling procedure[20]. To simulate the single-band 1D Hubbard model, we first prepared 1D systems by ramping up the large spacing component ($d_y = 2.3$ μm) of an optical superlattice in the y direction. The lattice was ramped linearly in two steps, first to $15E_r^y$ in 55 ms and then to $27E_r^y$ in 45 ms, which resulted in a final transverse tunnelling of $t_y/h = 1.2$ Hz. With a delay of 10 ms with respect to the start of the y-lattice ramp, the lattice along the tubes (x direction; spacing $d_x = 1.15$ μm) was turned on. The chosen ramp was again composed of two linear parts, the first was a ramp to $3E_r^x$ in 45 ms and the second to $5E_r^x$ in 55 ms. Simultaneously the scattering length was increased from $530a_B$ to $2{,}000a_B$ (where a_B is the Bohr radius) using a magnetic offset field close to the Feshbach resonance, which is located at 834.1 G. At the end of the ramps, the tunnelling along the Hubbard chains reached $t_x/h = 410$ Hz and the on-site interaction $U/h = 2.9$ kHz. The latter was calculated from the ground-band Wannier functions, neglecting higher-band corrections[34]. The corresponding final superexchange coupling was $J_x = 4t_x^2/U = h \times 235$ Hz.

To explore the Hubbard model in the 1D–2D crossover we first ramped up the large spacing component of the superlattice in the y direction to $0.2E_r^y$ in 60 ms and then to depths varying between $5E_r^y$ and $27E_r^y$ in 220 ms. The x-lattice ramp to depths varying between $9E_r^x$ and $10.6E_r^x$ in 280 ms started simultaneously with the second part of the y-lattice ramp. The magnetic offset field was adjusted to maintain a constant ratio of $U/t_x = 14$ at the end of the ramps. A local Stern–Gerlach detection technique[5] with a transverse magnetic field gradient of 95 G cm^{-1} was used to detect both the spin and the density on each lattice site with a fidelity of 97%.

Data analysis. Thanks to our local access to both the spin and the occupation at each lattice site in a single experimental run, we can group each Hubbard chain data by $\{j, N, S^z\}$, where j is the coordinate of the Hubbard chain in the y direction. This allows us to explore different filling and spin sectors (Extended Data Fig. 1). To study the effect of doping on spin correlations we only analysed the data in the $S^z = 0$ sector. The density profile along x is inhomogeneous and dependent on N and j, owing to the underlying harmonic confinement of $\omega = 2\pi \times 200(20)$ Hz. For each pair $\{j, N\}$ we computed a mean density profile $n_j(i, N)$ by averaging the occupation on each site i over different experimental realizations (Extended Data Fig. 2). The reported density, at which spin correlations between sites i and $i + x$ were analysed, is the mean density between the two points $n_j(i, x, N) = (1/x)\sum_{k=i}^{i+x} n_j(k, N)$

To highlight the oscillatory behaviour of the spin correlations as a function of density we considered the two-point spin correlations between sites i and $i+x$, conditioned on having single occupancies (filled circles) on these sites for each pair $\{j, N\}$:

$$C_{i,j,N}(x) = 4\langle S_i^z S_{i+x}^z \rangle_{\bullet_i \bullet_{i+x} j,N} + c(N) \tag{3}$$

where the angle brackets denote averaging over experimental runs, and $c(N)$ is a finite-size offset that depends on the atom number N and the temperature, which we experimentally found to be well described[21] by $c(N) = 1/(N-1) - 0.04(5)$. We also analysed the data in terms of the corresponding connected correlator and found them to agree with the non-connected version in equation (3) within statistical uncertainty. This check was also performed for all other non-connected correlators that we used in this work. Owing to the absence of density–density correlations beyond $x = 1$, this correlation function can be understood as being a renormalized two-point spin correlation[21] $C_{i,j,N}(x) \approx [4\langle S_i^z S_{i+x}^z \rangle_{i,j,N} - c(N)]/n_j(i, x, N)^2$. Finally we grouped all the $C_{i,j,N}(x)$ correlations according to their density $n_j(i, x, N)$ in bins of width $\Delta n = 0.1$ to compute the average spin correlation $C(x)$ for each n shown in Fig. 2.

The microscopic-scale origin of the incommensurate SDW is revealed by the spin correlations across holes and double occupancies shown in Extended Data Fig. 2b:

$$C_{i,y,N}^{\mathrm{dw}}(x) = 4\langle S_i^z S_{i+x}^z \rangle_{\bullet_i \circ_{i+1} \bullet_{i+x} y,N} + c(N)$$

where the open circles denote a doublon or a hole on site $i + 1$. The angle brackets indicate averaging over all experimental realizations in which these conditions are fulfilled. Both the holes and the doublons displace the spin correlations, leading to an increase of their wavelength.

To separate the effect of polarization on the spin correlations from the charge sector, we studied spin correlations in squeezed space. Here, we extend the concept of squeezed space to finite U by removing doublons and holes only when these are not nearest neighbours. The latter condition is supported by the strong doublon–hole bunching at $|x| = 1$, measured by $g_2(x) = -1 + \langle d_0 h_x \rangle / (\langle d_0 \rangle \langle h_x \rangle)$, where d_i and h_j denote the doublon and hole operators on sites i and j, respectively (see Extended Data Fig. 3a), which we attribute to quantum fluctuations. The full dataset in this measurement consisted of 26,442 Hubbard chains, which we prepared close to half-filling at the centre of the trap. This led to shorter chains with up to $N = 16$ atoms. We decided to use the squeezed space analysis instead of post-selecting the data to the zero-hole and -doublon sector to improve our statistics. Within statistical uncertainties, the post-selected data are consistent with these squeezed-state results.

The connected spin correlations in squeezed space, indicated by $\tilde{\imath}$ and $\tilde{x}$, are defined as

$$C_{\tilde{\imath},j,N_s,S^z}(\tilde{x}) = 4(\langle S_{\tilde{\imath}}^z S_{\tilde{\imath}+\tilde{x}}^z \rangle - \langle S_{\tilde{\imath}}^z \rangle \langle S_{\tilde{\imath}+\tilde{x}}^z \rangle) + c_{\mathrm{sq}}(N_s)$$

where the number of singly occupied sites N_s includes nearest-neighbour doublon–hole pairs. We again take into account a finite-size offset $c_{\mathrm{sq}}(N_s)$ similar[21] to $c(N)$, which we obtain by taking the mean of the correlator at distances $\tilde{x} = 4, \ldots, 7$. The magnetization of the effective Heisenberg chain in squeezed space is defined as $m = S^z/N_s$. We group all the $C_{\tilde{\imath},j,N_s,S^z}(\tilde{x})$ correlations by their polarization m in bins of width $\Delta m = 0.04$ to compute the average spin correlation C shown in Fig. 3. The Fourier transform shown in Fig. 3b is calculated using our experimental points $\tilde{x} = 1, \ldots, 10$, which limits our $\tilde{k}$-space resolution, and is zero-padded to increase the visibility of the SDW vector change.We highlight in Extended Data Fig. 4 the quantitative agreement in the fixed-distance comparison of $C(\tilde{x})$ with exact diagonalization results of the Heisenberg chain at $T = 0.7J$ averaged over our experimental $\{S^z, N_s\}$ distribution. This validates the use of the squeezed-space concept away from $S^z = 0$. Similarly to the doped case, the microscopic-scale origin of the polarization dependence of the SDW wavelength can be explained in terms of domain walls in squeezed space. From the sum rule for conditional probabilities we can write

$$\begin{aligned} C(\tilde{x}) = \Big(& 4\langle S_{\tilde{\imath}}^z S_{\tilde{\imath}+\tilde{x}}^z \rangle_{S^z \sigma_{\tilde{\imath}+1} > 0} \times p_{S^z \sigma_{\tilde{\imath}+1} > 0} \\ & + 4\langle S_{\tilde{\imath}}^z S_{\tilde{\imath}+\tilde{x}}^z \rangle_{S^z \sigma_{\tilde{\imath}+1} < 0} \times p_{S^z \sigma_{\tilde{\imath}+1} < 0} \Big) \\ & - 4\langle S_{\tilde{\imath}}^z \rangle \langle S_{\tilde{\imath}+\tilde{x}}^z \rangle \end{aligned} \tag{4}$$

where $\langle X \rangle_Y$ denotes the conditional expectation value of X given Y, and p_Y is the probability of event Y. Intuitively, for low enough polarization, the spin correlations at short distances around the minority spins will remain in phase with the unpolarized case on average to minimize the magnetic energy. The short-distance spin correlations around the majority spins, on the other hand, are expected to be more sensitive to polarization. We consider the oscillating part of the two conditional expectations in the parentheses in equation (4) for $m = -0.12$ in Fig. 3d after removing a finite-size and polarization-dependent offset similar to $c_{\mathrm{sq}}(N_s)$. We indeed find for $\tilde{x} \leq 5$ antiferromagnetic correlations in phase with the unpolarized case across the minority spins, whereas these are phase shifted by π around the majority spins. The sign change of the spin correlations across the majority spins reveals that the polarization is mostly carried by pairs of parallel majority spins, which we find to be delocalized over the full chain. We directly detected pairs of parallel spins, computed the spin correlations across them (inset of Fig. 3d) and found that they indeed act as delocalized domain walls in the antiferromagnet. These domain walls stretch spin correlations and lead to incommensurate magnetism when $m \neq 0$. Examples of domain-wall distributions are shown in Extended Data Fig. 5.

In the dimensional crossover and 2D regime we prepared anisotropic samples consisting of about five coupled Hubbard chains (Extended Data Fig. 6). Similarly to the 1D case, the spin correlations were calculated on singly occupied sites through

$$C_{\boldsymbol{n}}(\boldsymbol{r}) = 4\langle S_{\boldsymbol{n}}^z S_{\boldsymbol{n}+\boldsymbol{r}}^z \rangle_{\bullet_{\boldsymbol{n}} \bullet_{\boldsymbol{n}+\boldsymbol{r}}}$$

averaged over all sites $\boldsymbol{n} = (n_x, n_y)$ where n_i are integers labelling lattice sites. When studying spin correlations around double occupancies and holes at the crossover, we minimized biasing of the correlator by a possibly distorted magnetic background around quantum-fluctuation-induced doublon–hole pairs. The strong bunching observed in the doublon–hole correlations $g_2(\boldsymbol{r})$ (see Extended Data Fig. 6) at the nearest-neighbour scale identifies a strong contribution of quantum fluctuations to these correlations. Hence, we discarded any doublons with one unoccupied nearest neighbour from the analysis of the spin correlations.

© 2019 Springer Nature Limited. All rights reserved.

Quantum Monte Carlo calculations. The QMC results reported here were obtained in a similar fashion as those found in ref. [5]. Simulating the fermionic system was possible by mapping between the 1D fermionic Hubbard model and a system of two hard-core bosonic species with on-site interspecies interactions[35].

We used the worm algorithm[36] in the implementation of ref. [37]. This algorithm exhibits a linear scaling in the system volume when simulating the resulting bosonic model. The spin S_i at site i of the fermionic model is mapped onto a diagonal observable with respect to the Fock basis $\{|\ldots, n_j, \ldots\rangle\}$ of the bosonic model, which is proportional to the differences in the occupation numbers of the bosonic particles at the same site.

The simulations were all carried out in the grand canonical ensemble. The system consisted of a homogeneous lattice of $L = 20$ sites with hard-wall boundary conditions. This size was confirmed to be large enough to avoid finite-size corrections. We note however that the correlations were affected by an unavoidable systematic offset that scales as $1/N$, which was corrected for in the analysis, as explained above.

To better mimic the measurement procedure of the actual experiment, we saved the raw QMC configurations and performed the analysis off-line. In this process, care must be taken to make sure that subsequent configurations are decorrelated. A further blocking and jackknife estimation was used to rule out any residual correlation.

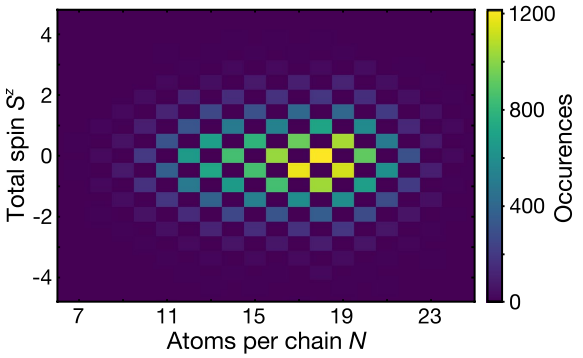
The off-line analyses were subject to the same filtering procedures for the occupation and magnetization sector as in the experimental procedure. To gather enough statistics for the different values of the density that were accessible in the experiment, we tuned the chemical potentials of the two bosonic species so as to have symmetric mixtures with total density n between 0.4 and 1.2.

Data availability

The datasets generated and analysed during this study are available from the corresponding author upon reasonable request.

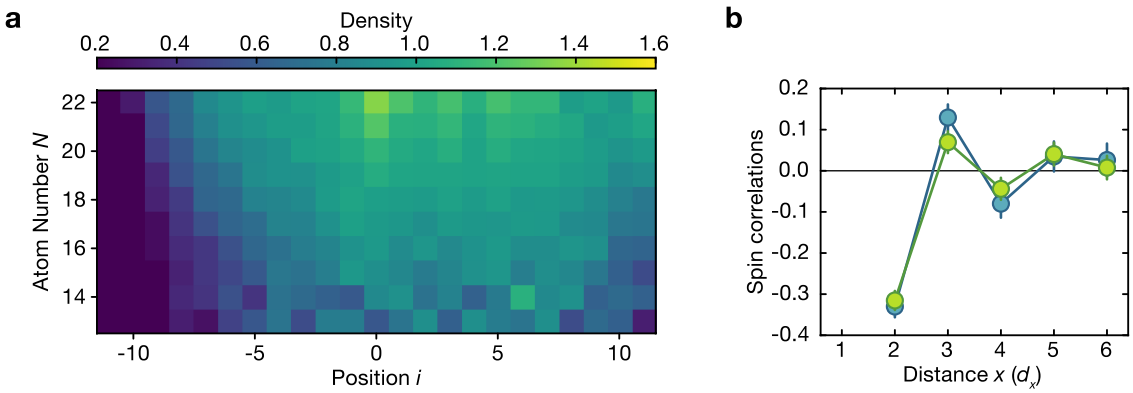
34. Büchler, H. P. Microscopic derivation of Hubbard parameters for cold atomic gases. *Phys. Rev. Lett.* **104**, 090402 (2010).
35. Jordan, P. & Wigner, E. Über das Paulische Äquivalenzverbot. *Z. Phys.* **47**, 631–651 (1928).
36. Prokof'ev, N. V., Svistunov, B. V. & Tupitsyn, I. S. Exact, complete, and universal continuous-time worldline Monte Carlo approach to the statistics of discrete quantum systems. *J. Exp. Theor. Phys.* **87**, 310–321 (1998).
37. Pollet, L., Houcke, K. V. & Rombouts, S. M. Engineering local optimality in quantum Monte Carlo algorithms. *J. Comput. Phys.* **225**, 2249–2266 (2007).

© 2019 Springer Nature Limited. All rights reserved.



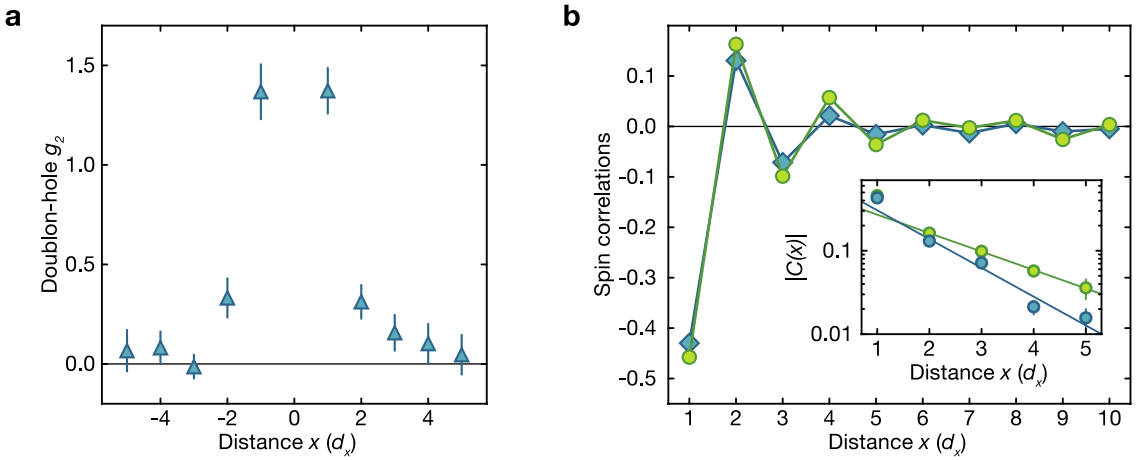
Extended Data Fig. 1 | Chain statistics. Hubbard chain statistics are shown for a typical dataset containing 5,240 shots. The total spin S^z and total atom number N of individual Hubbard chains are conserved quantities of the Hamiltonian for each experimental run. However, they fluctuate for different experimental realizations, allowing us to explore the effects of doping and polarization individually through data grouping.

© 2019 Springer Nature Limited. All rights reserved.



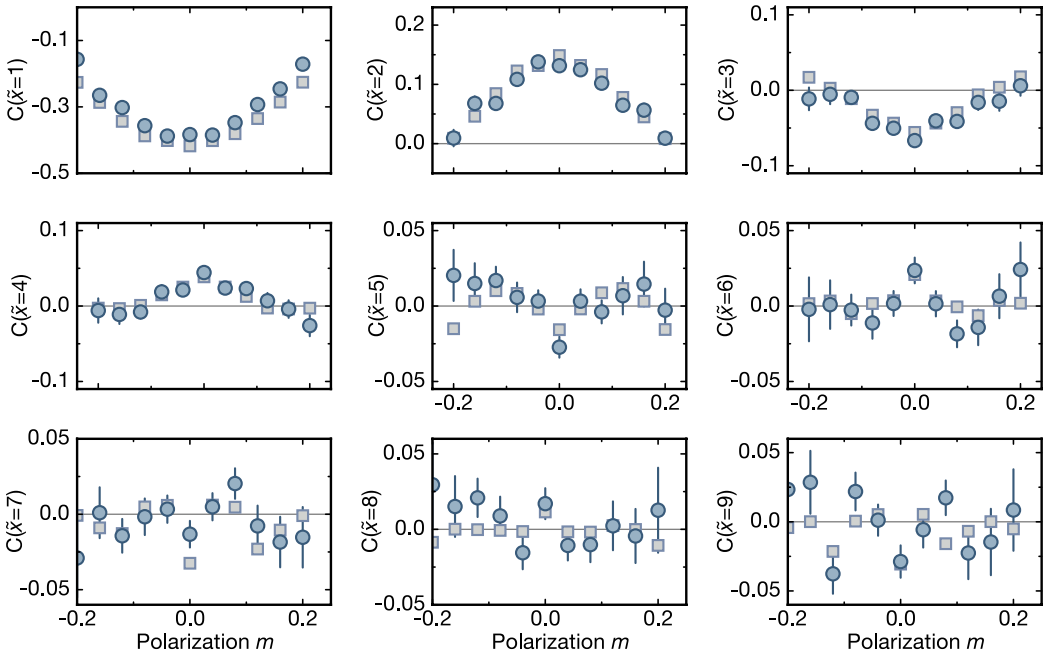
Extended Data Fig. 2 | Density properties of the 1D clouds. a, Density profiles $n_0(i, N)$ of the chain located at the centre of the cloud in the y direction ($j = 0$). **b**, Antiferromagnetic spin correlations across a hole fixed at $x = 1$ (green) or a doublon (blue), measured through $C^{\mathrm{dw}}(x)$. The correlation signal is shifted by the hole, which is the microscopic-scale origin of the incommensurate spin correlations away from half-filling in the spin-balanced case. Error bars denote one standard error of the mean.

© 2019 Springer Nature Limited. All rights reserved.



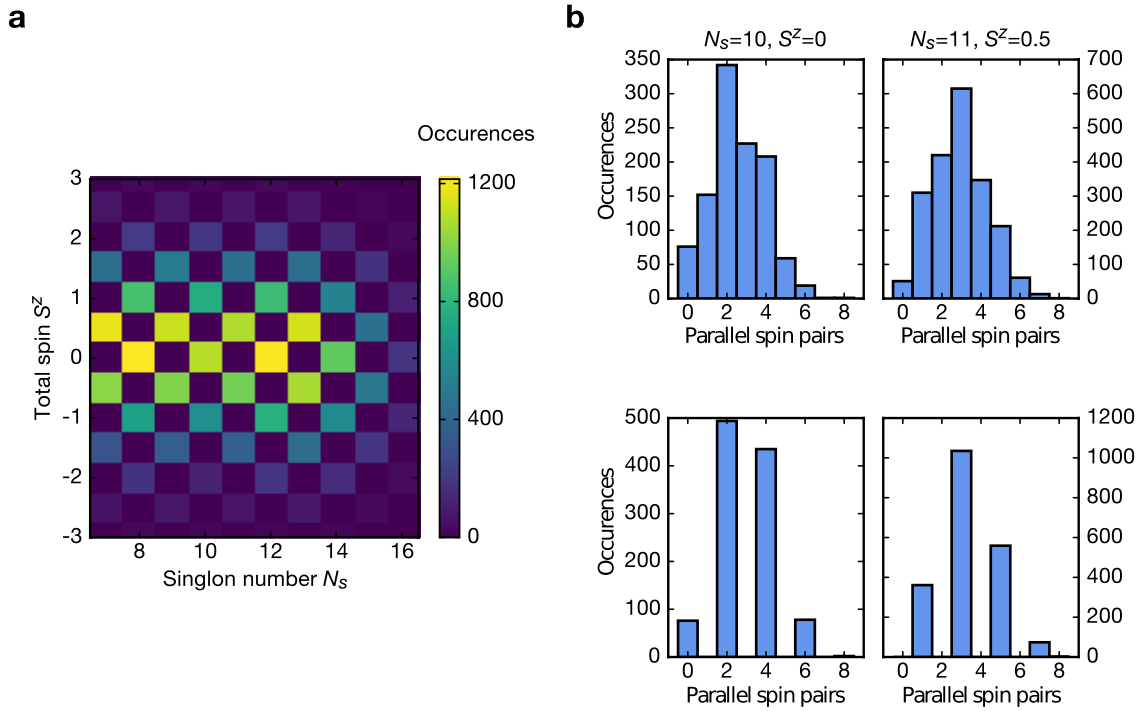
Extended Data Fig. 3 | Spin correlations in squeezed space. a, Doublon–hole correlations, measured by $g_2(x)$. The strong bunching at $|x| = 1$ reveals neighbouring doublon–hole pairs as mostly stemming from quantum fluctuations. This justifies our extension of the squeezed-space concept away from $U \to \infty$. **b**, Spin correlations in the zero-magnetization sector at the centre of the cloud. Averaging over different polarizations (blue) results in a faster decay of the spin correlations with distance x in squeezed space compared to the $S^z = 0$ sector (green). Exponential fits of the correlation envelope for distances $x = 2, \ldots, 6$ yield $\xi_{\mathrm{avg}} = 1.3(1)$ without magnetization post-selection and $\xi_0 = 2(1)$ in the $S^z = 0$ sector. Error bars denote one standard error of the mean.

© 2019 Springer Nature Limited. All rights reserved.



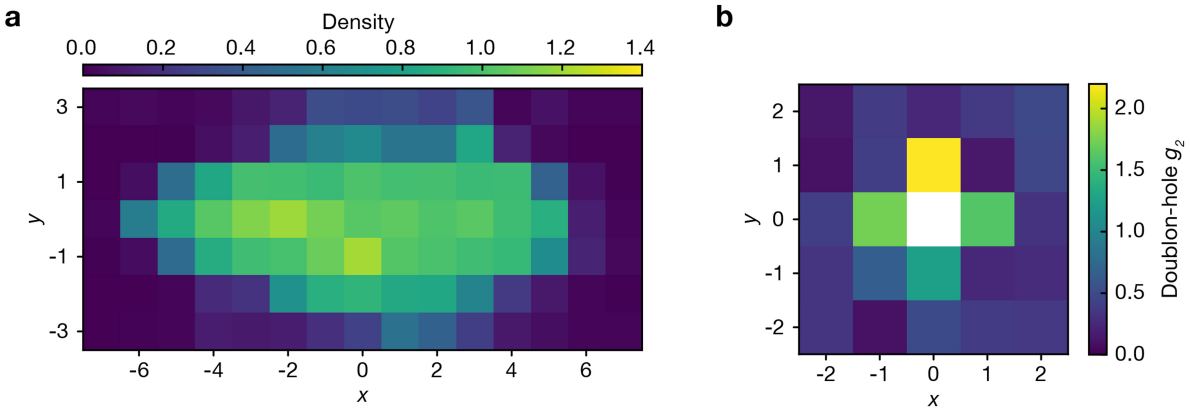
Extended Data Fig. 4 | Squeezed-space spin correlations at fixed distance. Experimental spin correlations in squeezed space are shown as a function of polarization m for distances $\tilde{x} = 1, \ldots, 9$ (blue circles), along with exact diagonalization results for spin correlations in the Heisenberg chain at $T = 0.7J$ averaged over the experimental $\{S^z, N_s\}$ distribution (grey squares). Error bars denote one standard error of the mean.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Chain statistics for the polarization study. **a**, Experimental distribution $\{S^z, N_s\}$ used for studying the effects of polarization on the SDW vector. **b**, Histograms of pairs of parallel spins for $\{N_s = 10, S^z = 0\}$ and $\{N_s = 11, |S^z| = 0.5\}$. The upper row shows, as expected, an upward shift of the distribution towards larger number of domain walls away from $S^z = 0$. By using the convention that spins pointing in the same direction at the edges contribute as one pair of parallel spins (lower row), we find that the parity of the number of domain walls is even in the integer-spin sectors and odd in the half-integer case. In the $S^z = 0$ sector, domain walls appear in pairs of opposite quantum numbers, which do not affect the SDW wavevector. In the $|S^z| = 0.5$ case on the other hand, we find a minimum of one domain wall owing to the excess spin and higher numbers of domain walls corresponding to pairs of additional excited parallel-spin pairs with opposite quantum numbers.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Properties of the prepared 2D clouds. a, Density distribution for $t_y/t_x = 1$. **b**, Doublon–hole correlations $g_2(\boldsymbol{r})$. The strong bunching of the doubon–hole correlations $g_2(\boldsymbol{r})$ at $|\boldsymbol{r}| = 1$ justifies the rejection of outcomes in which holes and doublons are found nearby when studying the effects of doping.

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0809-4

An ultrafast symmetry switch in a Weyl semimetal

Edbert J. Sie[1,2,13], Clara M. Nyby[3,13], C. D. Pemmaraju[2], Su Ji Park[2], Xiaozhe Shen[4], Jie Yang[4,5], Matthias C. Hoffmann[4], B. K. Ofori-Okai[4,6], Renkai Li[4], Alexander H. Reid[4], Stephen Weathersby[4], Ehren Mannebach[7], Nathan Finney[8], Daniel Rhodes[9,10,12], Daniel Chenet[8], Abhinandan Antony[8], Luis Balicas[9,10], James Hone[8], Thomas P. Devereaux[1,2], Tony F. Heinz[2,5,11], Xijie Wang[4] & Aaron M. Lindenberg[2,5,7]*

Topological quantum materials exhibit fascinating properties[1–3], with important applications for dissipationless electronics and fault-tolerant quantum computers[4,5]. Manipulating the topological invariants in these materials would allow the development of topological switching applications analogous to switching of transistors[6]. Lattice strain provides the most natural means of tuning these topological invariants because it directly modifies the electron–ion interactions and potentially alters the underlying crystalline symmetry on which the topological properties depend[7–9]. However, conventional means of applying strain through heteroepitaxial lattice mismatch[10] and dislocations[11] are not extendable to controllable time-varying protocols, which are required in transistors. Integration into a functional device requires the ability to go beyond the robust, topologically protected properties of materials and to manipulate the topology at high speeds. Here we use crystallographic measurements by relativistic electron diffraction to demonstrate that terahertz light pulses can be used to induce terahertz-frequency interlayer shear strain with large strain amplitude in the Weyl semimetal WTe_2, leading to a topologically distinct metastable phase. Separate nonlinear optical measurements indicate that this transition is associated with a symmetry change to a centrosymmetric, topologically trivial phase. We further show that such shear strain provides an ultrafast, energy-efficient way of inducing robust, well separated Weyl points or of annihilating all Weyl points of opposite chirality. This work demonstrates possibilities for ultrafast manipulation of the topological properties of solids and for the development of a topological switch operating at terahertz frequencies.

Topological materials provide a platform from which to pursue exotic physics from the seemingly distant field of particle physics in condensed matter systems, such as Weyl fermions[12–14]. This has led to a worldwide effort focused on discovering new topological quantum materials. One prime example is WTe_2, which is a layered transition-metal dichalcogenide (TMD) that crystallizes in a distorted hexagonal net with an orthorhombic unit cell, where the W–W chain direction is along the crystallographic axis a (Fig. 1a)[15]. The lack of inversion symmetry in this material leads to a topological semimetal (predicted[16] and experimentally verified[17–19]) with type II Weyl points (WPs), which can be manipulated through atomic-scale lattice distortions. In our experiments, the measured shear displacement amplitudes (about 1%) are more than sufficient to undergo a complete annihilation of the WPs or a more-than-twofold increase in WP separation, depending on the direction of the shear displacement. Ordinary crystals start to fracture at these strains, and conventional means of using piezoelectric transducers to induce lattice distortions typically reach about 0.05% strain[20]. The observed large strain in our experiment (about 1%) is possible because we use light to generate interlayer shear strain in a weakly van-der-Waals-bonded two-dimensional TMD—a method that is less susceptible to lattice damage than uniaxial straining but that can considerably alter the electronic band structure.

We used a relativistic ultrafast electron diffraction (UED) technique to reconstruct the shear motion and crystallographically quantify the corresponding atomic displacements by measuring more than 200 Bragg peaks (Fig. 1b)[21] (Methods). We used two different terahertz pump excitation schemes, involving a quasi-single-cycle excitation at 3 THz and a few-cycle excitation at 23 THz, both of which enable application of an all-optical bias field while minimizing interband transitions (Methods). The arrival time of the electron beam (probe) can be adjusted with respect to the terahertz pulses (pump) using an optical delay stage. The measured diffraction pattern in equilibrium (Fig. 1c) is consistent with the orthorhombic phase of WTe_2 (Fig. 1a). By applying terahertz pump pulses, we find that the intensities of many Bragg peaks are modulated, which indicates structural changes in the WTe_2 lattice. To investigate the lattice dynamics, we plot the intensity changes $\Delta I/I_0$ of several Bragg peaks as a function of the time delay, Δt, between the terahertz pump and electron probe pulses (Fig. 1d, e). The time traces show coherent oscillations with a frequency of 0.24 THz (Fig. 1f), which is consistent with a low-frequency interlayer shear phonon mode predicted by density functional theory (DFT) analysis (Methods).

To determine the atomic motion, we plot the measured $\Delta I/I_0$ image at 2.5 ps (Fig. 2a). We find that the changes in peak intensity alternate along the b axis, while peaks (hkl) with $k = 0$ show negligible changes. This suggests that the interlayer shear displacement is along the b axis. To verify this, we derive the peak intensity modulation $\Delta I/I_0$ by introducing a top-layer shear displacement Δy with respect to the bottom layer into the structure factor

$$S(\Delta y) = 2\sum_{\text{top}} f_j \cos[2\pi(hx_j + ky_j) + \pi k\Delta y] \quad (1)$$

where we have used the underlying crystalline symmetry in WTe_2 to obtain the final, simplified expression (Methods). Here, the summation runs over all atoms in the top half of the unit cell (two W and four Te atoms), f_j is the atomic scattering factor, $(x_j\hat{\boldsymbol{a}}, y_j\hat{\boldsymbol{b}}, z_j\hat{\boldsymbol{c}})$ is the atom position in the unit cell, $hk0$ are the usual Miller indices for the [001] zone axis, and the lattice constants are $a = 3.477$ Å, $b = 6.249$ Å and $c = 14.018$ Å. To compare this with our experiment, we calculate the peak intensity modulation $\Delta I \propto |S(\Delta y)|^2 - |S(0)|^2$, shown in Fig. 2b. Here, we define a positive shear displacement ($\Delta y > 0$), as shown in the inset of Fig. 2e. The simulated image shows excellent agreement with the measured image and verifies that the peak intensity modulation arises primarily from interlayer shear displacement along the b axis.

We determine the shear displacement amplitude through a global fitting between the simulated and measured intensities of many Bragg

[1]Geballe Laboratory for Advanced Materials, Stanford University, Stanford, CA, USA. [2]SIMES, SLAC National Accelerator Laboratory, Menlo Park, CA, USA. [3]Department of Chemistry, Stanford University, Stanford, CA, USA. [4]SLAC National Accelerator Laboratory, Menlo Park, CA, USA. [5]PULSE Institute, SLAC National Accelerator Laboratory, Menlo Park, CA, USA. [6]Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA, USA. [7]Department of Materials Science and Engineering, Stanford University, Stanford, CA, USA. [8]Department of Mechanical Engineering, Columbia University, New York, NY, USA. [9]National High Magnetic Field Laboratory, Florida State University, Tallahassee, FL, USA. [10]Department of Physics, Florida State University, Tallahassee, FL, USA. [11]Department of Applied Physics, Stanford University, Stanford, CA, USA. [12]Present address: Department of Mechanical Engineering, Columbia University, New York, NY, USA. [13]These authors contributed equally: Edbert J. Sie, Clara M. Nyby. *e-mail: aaronl@stanford.edu

© 2019 Springer Nature Limited. All rights reserved.

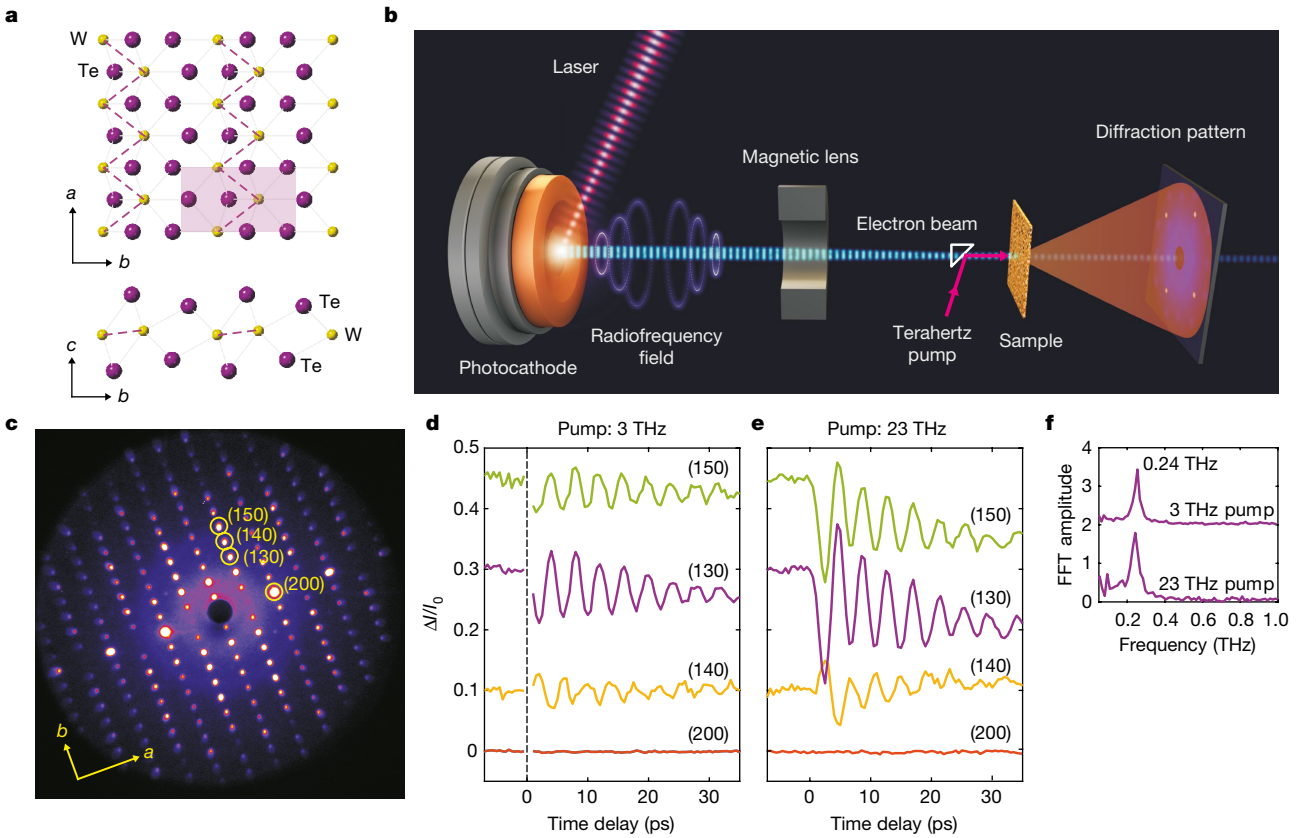


Fig. 1 | Observation of coherent interlayer shear displacements in WTe_2 measured using relativistic ultrafast electron diffraction. a, Lattice structure of Td WTe_2: top view (*a*–*b* plane) and side view (*b*–*c* plane). The dashed lines indicate the W–W zigzag chain along the *a* axis. The shaded area shows the unit cell. **b**, Schematic of SLAC 3-MeV relativistic ultrafast electron diffraction setup. Image courtesy of G. Stewart. The electron beam is generated using ultraviolet femtosecond laser pulses at the photocathode and accelerated using an intense radiofrequency field from the klystron. The diffracted electron beam, measured using an EMCCD (electron-multiplying charge-coupled device) camera, is used to probe the structural changes of the sample. Intense terahertz pump pulses are used to induce interlayer shear strain in WTe_2. **c**, Measured diffraction pattern of WTe_2 at equilibrium. **d**, **e**, Changes in Bragg peak intensity as a function of time delay between the terahertz pump pulses and the electron beam. **f**, Fast Fourier transform (FFT) amplitude of the oscillations, indicating the 0.24-THz shear phonon mode along the *b* axis. Curves in **d**–**f** are offset for clarity.

peaks as a function of time delay (Methods). The fitting results (Fig. 2c, d) show that the proposed interlayer shear displacement fits the experimental data very well. At a low pump field of 2.6 MV cm^{-1} (23 THz), the fitting results (Fig. 2f; red line) show shear displacements that oscillate between −1.7 to +3.6 pm in the early stage and gradually develop an offset towards a positive shear displacement of +1.5 pm in the later stage. Increasing the pump field to 7.5 MV cm^{-1} (Fig. 2f; blue line) leads to a much larger offset of +8.0 pm (1.3% strain) that gradually builds up on a timescale of 25 ps and persists for longer than 70 ps. This long-lived offset indicates that the lattice has found a new equilibrium position, deviating from the simple harmonic oscillator behaviour that is normally expected. The shear oscillation frequency decreases at increasing field strengths (Extended Data Fig. 6), consistent with an anharmonic response as the material is driven to large amplitudes.

To investigate the driving mechanism of this behaviour, we measured the shear amplitude as a function of pump field strength and polarization. We found that the amplitude increases linearly with the field under different off-resonance frequencies (Fig. 3a) and is polarization-isotropic (Fig. 3b, c). Moreover, the shear motion always starts towards positive displacement regardless of polarization (Fig. 3b). This behaviour cannot be explained through the infrared active and Raman (impulsive stimulated Raman scattering) mechanisms normally considered. We propose a terahertz-field-driven charge-current mechanism, as indicated by the linear amplitude response to field strength and motivated by recent calculations[22] that predict a transition from an orthorhombic (Td) to a monoclinic and centrosymmetric (1T′) phase in WTe_2 via hole doping at a density of about 10^{20} cm^{-3}. Microscopically, the applied field accelerates the electron population away from the topmost valence band, which constitutes an interlayer antibonding orbital. This destabilizes the interlayer coupling strength and launches a shear motion along the in-plane transition pathway from the Td to the 1T′ phase with a new equilibrium position ($\Delta y > 0$) (Fig. 2e, Extended Data Fig. 1). In our experiment, the effective hole doping density can be estimated using the Drude model, which gives a doping density of about 10^{20} cm^{-3}, comparable to the impulsive driving force for the interlayer shear motion (Methods).

We note that a departure from the Td phase via interlayer displacement could result in two possible phases, monoclinic 1T′ and orthorhombic 1T′(*), both of which are centrosymmetric. Unlike the 1T′ phase, the 1T′(*) phase can be reached while maintaining the orthorhombic structure (Extended Data Fig. 1). Although the observed long-lived offset (8 pm; Fig. 2f) is somewhat smaller than that calculated for a complete transition to the 1T′(*) phase (about 10–15 pm; Fig. 2e), the measured displacement should be viewed as a lower bound due to spatial averaging of the film seen by the electron beam. This is because under the simplistic approximation that the induced metastable phase is a centrosymmetric phase, either 1T′(*) or 1T′, the volume change associated with this transformation results in a complex longitudinally heterogeneous strain profile with strain waves propagating from the interfaces[23] and complicated by substrate interactions. These processes probably underlie some of the complex longer-timescale dynamics (of the order of 25 ps) shown in Fig. 2f. This timescale is consistent with the shear-wave propagation time across the sample thickness (Methods; Extended Data Fig. 4), which is longer than recently

© 2019 Springer Nature Limited. All rights reserved.

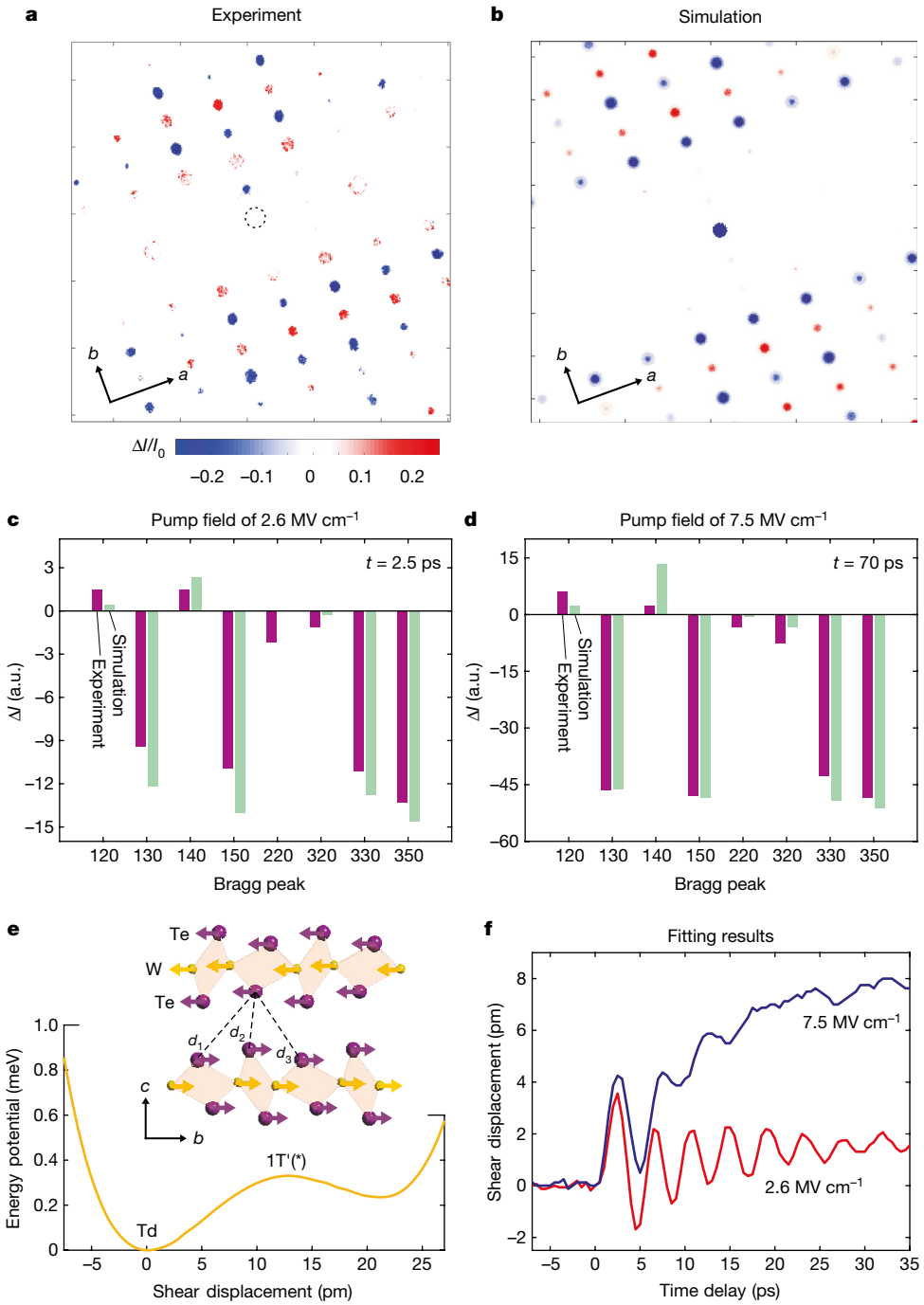


Fig. 2 | Determination of interlayer shear atomic displacements in WTe_2. **a**, Measured diffraction intensity changes at 2.5 ps, obtained using a pump frequency of 23 THz. The alternating sign changes along the *b* axis are signatures of shear displacements along the *b* axis, as shown in the inset of **e**. **b**, Simulated diffraction intensity changes, obtained by rigidly displacing the adjacent WTe_2 layers relative to each other (shear) along the *b* axis. **c**, **d**, Bar charts showing the ΔI fitting results between the experiment and simulation to obtain Δy at a pump field of 2.6 MV cm^{-1} ($\Delta t = 2.5$ ps; **c**) and 7.5 MV cm^{-1} ($\Delta t = 70$ ps; **d**); a.u., arbitrary units. **e**, Energy potential as a function of interlayer shear displacement (see DFT analysis, Methods). The schematic shows an interlayer shear motion along the positive displacement, that is, towards $d_1 < d_3$. The Td phase is the ground-state non-centrosymmetric structure, whereas the 1T′(*) phase is an excited-state centrosymmetric structure that is accessible via an interlayer shear displacement. **f**, Shear displacements as a function of time delay at pump fields of 2.6 MV cm^{-1} (red) and 7.5 MV cm^{-1} (blue), obtained through global fitting of the intensity changes of many Bragg peaks.

observed photoinduced structural transitions in atomically thin indium wires[24]. To further understand this, we carried out comparative measurements in a related material, $MoTe_2$, which can crystallize in the Td and 1T′ phases at different temperatures[25,26]. Whereas in the Td phase we observe light-induced shear displacements, in the 1T′ phase we observe negligible signals (Methods). This is consistent with a mechanism in which the terahertz fields drive the material unidirectionally towards 1T′.

The ability to drive a shear displacement using terahertz pulses offers a way to manipulate the topological properties of the semimetal WTe_2 on ultrafast timescales. There are a total of eight WPs in the equilibrium Td phase of WTe_2 in the $k_z = 0$ plane. It is sufficient to consider two of these WPs in the $k_x, k_y > 0$ quadrant because we can obtain the remaining six WPs through the time-reversal and mirror symmetries. The two WPs carry opposite chiralities associated with the topological charges $\chi^- = -1$ (WP1) and $\chi^+ = +1$ (WP2), which are connected by a Fermi

© 2019 Springer Nature Limited. All rights reserved.

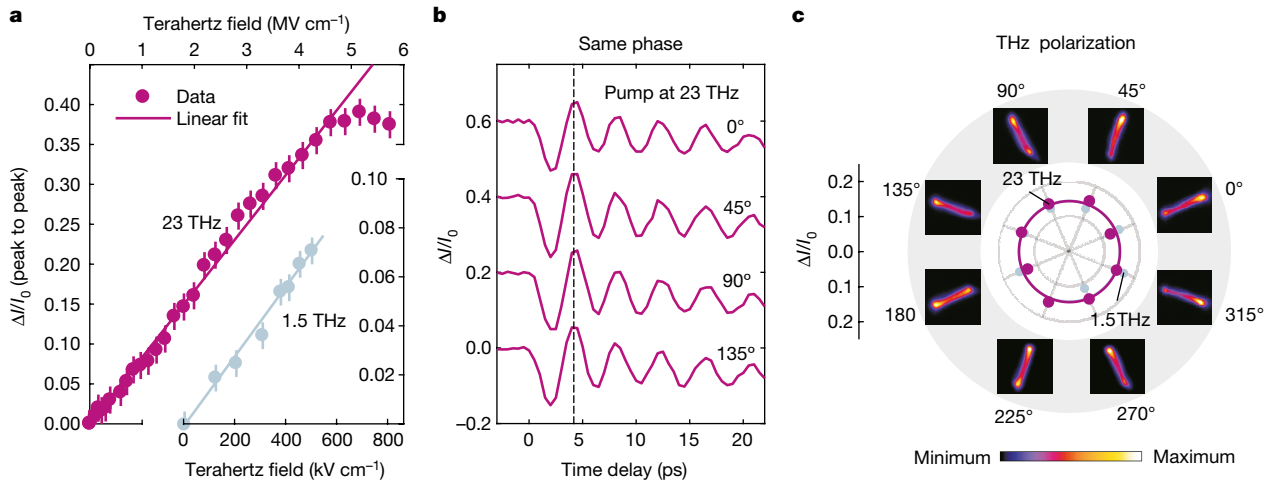


Fig. 3 | Field and polarization dependence of terahertz-induced shear amplitudes. a, Bragg peak (130) intensity changes at increasing terahertz-field strength at frequencies of 1.5 THz and 23 THz. Both experiments show a linear dependence with terahertz electric field. The vertical error bars represent standard deviations. **b**, Time trace at various terahertz polarizations (23-THz pump), showing that the modulation always starts at the same phase and amplitude, regardless of the pump polarization. A similar feature is also observed using a 1.5-THz pump. Curves in **b** are offset for clarity. **c**, Polar plot of the oscillation amplitude at various pump polarizations using frequencies of 1.5 THz (grey circles) and 23 THz (purple circles). The shaded areas show the EMCCD images of the electron-beam transverse displacement by the terahertz field (streaking), as a measure of the terahertz-pump polarization. These results show that the driving mechanism of the shear mode is linear in the applied field strength and isotropic in the polarization.

arc on the surface. Because the two WPs are separated mainly along k_y, we can expect large changes in the WP separation in momentum space by tuning the hopping parameters and band dispersion through interlayer shear strain along the y axis. In this way, the induced shear strain acts on the Weyl fermions as a chiral gauge-field vector potential, $\boldsymbol{A}$, because it couples to WP1 and WP2 with opposite signs in momentum space, $\boldsymbol{p} \rightarrow (\boldsymbol{p} - \chi^{\pm} e\boldsymbol{A})$, where e is the electron charge.

To demonstrate this, we compute the electronic band structure of WTe_2 using first-principles DFT calculations and monitor the positions of the WPs at different interlayer displacements Δy (Fig. 4a), as determined by our UED results. Our DFT calculations are performed using a Born–Oppenheimer approximation, in which electrons can instantaneously adjust to the new lattice environment. This is particularly appropriate for the interlayer shear mode because its timescale is much longer than that of the electron's, and the use of a terahertz pump does not create substantial electronic excitation. Type II WPs result from crossings between electron and hole bands. Hence, by mapping the energy difference between the two bands in momentum space, the positions of the WPs can be identified as the zero-energy gap position (WP1, blue; WP2, red). At equilibrium ($\Delta y = 0$), the WPs are separated by 0.7% of the reciprocal lattice vector $|\boldsymbol{G}_2|$. At positive shear displacements, the WPs move towards each other and mutual annihilation occurs at $\Delta y = +2.2$ pm. At negative shear displacements, the WPs move away from each other, leading to a more robust topological structure with more than twofold-increased WP separation. This is consistent with the intuitive picture in which positive shear moves towards a centrosymmetric and trivial phase, whereas negative shear moves towards a non-centrosymmetric and topological phase. At increasingly negative Δy, WP2 approaches the mirror plane at $k_y = 0$ and eventually annihilates with its mirror image of opposite chirality. This leads to a Lifshitz transition from a topological semimetal with eight WPs to one with four WPs, achieving the minimum non-zero number of WPs allowed in a time-reversal invariant system (Fig. 4b).

Although it is challenging to measure the distinct topological phases across the Lifshitz transition, we can experimentally verify the transition from a topological phase to a trivial phase using a time-resolved second-harmonic generation (SHG) technique. In a situation where inversion symmetry in WTe_2 is restored, the electronic phase transition from a topological to a trivial semimetal must follow. This is because the emergence of WP pairs in materials is contingent on lifting the double degeneracy of a Dirac cone by either breaking time-reversal or inversion symmetry. SHG arises from a non-zero second-order susceptibility, as shown in non-centrosymmetric topological systems[27,28]. Thus, it can be used as a sensitive probe to monitor the inversion symmetry and topological changes in WTe_2. In this measurement, we use a 2.1-μm-wavelength pump pulse to induce the transition, which gives interlayer shear displacements similar to those induced by terahertz pulses. Figure 4d shows the SHG polarization scans of WTe_2 in the absence of pump pulse (blue), which shows two lobes oriented horizontally. After the pump pulse arrives ($\Delta t = 2$ ps), the SHG vanishes almost completely in all polarizations, with magnitudes comparable to the detection noise level and to the measured SHG signal from centrosymmetric 1T′ $MoTe_2$. Figure 4e shows the measured SHG time trace from WTe_2 at various pump field strengths. At low fields (blue and red curves), the SHG intensity oscillates with the frequency of the shear mode at 0.24 THz. Of particular importance is that the oscillation always starts with a reduced SHG towards a centrosymmetric structure, consistent with our UED results (Fig. 2e, f). At high field (yellow) the SHG intensity plummets drastically, approaching a 100% reduction, which persists beyond the nanosecond timescale (Extended Data Fig. 7c). This indicates that WTe_2 exhibits a phase transition from a non-centrosymmetric to a centrosymmetric phase, consistent with our diffraction studies, and must correspond to a topological-to-trivial phase transition.

Similar manipulations of the WPs in WTe_2 can in principle be obtained through a compressive uniaxial strain along the a axis[16]. For example, at 1% uniaxial strain the WP separation is 2.2% of $|\boldsymbol{G}_2|$, and annihilation of WP2 at the mirror plane occurs at 2% uniaxial strain with energy cost of 32–39 meV per unit cell. This is about 1–2 orders of magnitude larger than the energy required for a shear strain to cause the same effect (Methods), indicating that the interlayer strain provides a more energy-efficient means of manipulating the topological band structure. In addition, shear displacement allows manipulation of WPs at terahertz frequencies. This ultrafast motion of the WPs is associated with a time-varying elastic gauge potential $A(t)$ and yields a pseudoelectric field $E = -\partial A/\partial t$, which can be used as a means of modulating charge density between bulk and surface[9] and to explore effective gauge fields driven by space- and time-dependent strains[8,9,29,30].

These findings offer a new promising way to enhance control over the topological properties of matter by modulating the topological invariants through field-driven lattice deformations. This leads to a substantial motion of the WPs and an ultrafast switch to structures with

© 2019 Springer Nature Limited. All rights reserved.

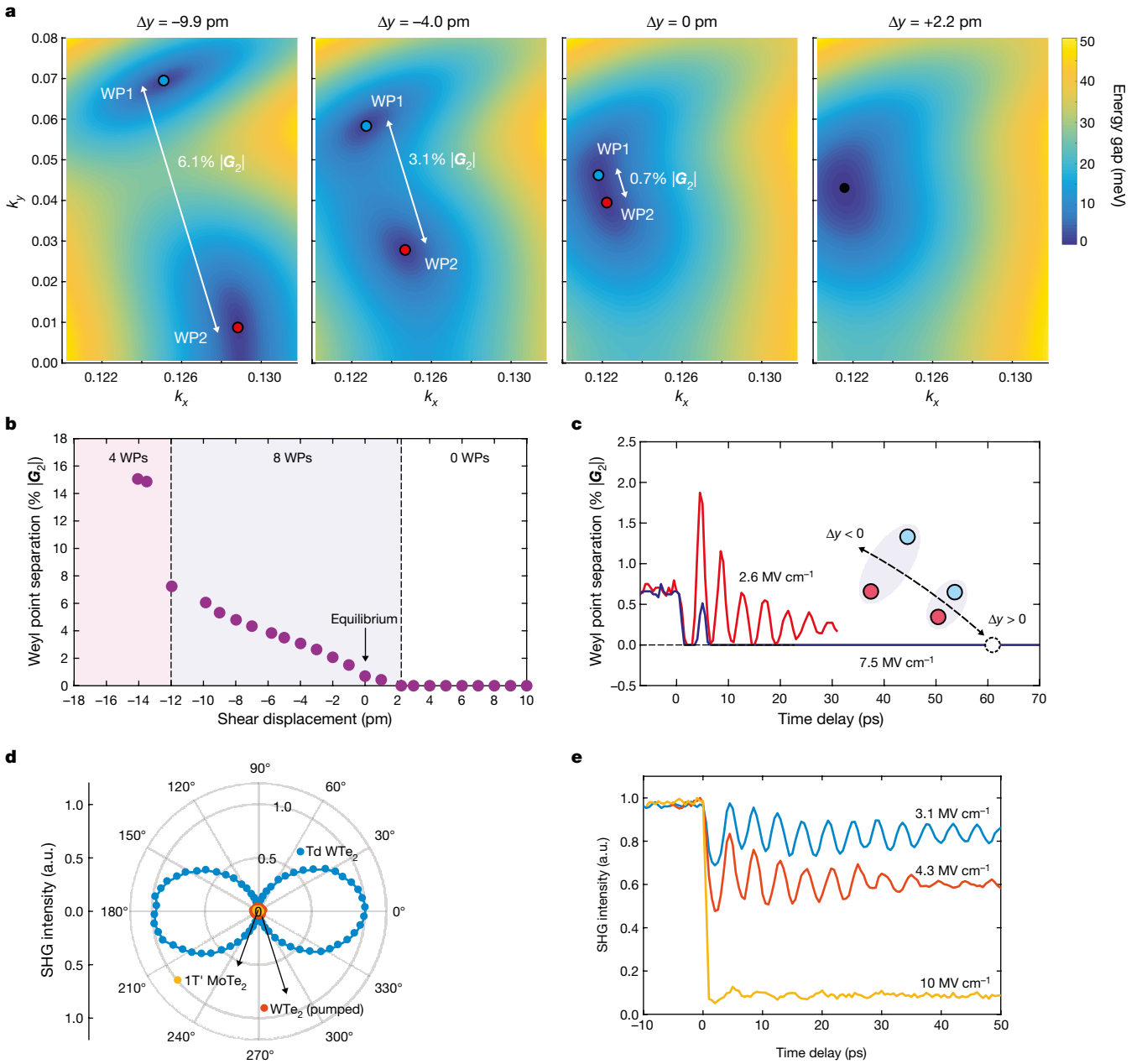


Fig. 4 | Strain-induced WP separation and topological phase transition. **a**, The two nearest WPs in momentum space (WP1, blue; WP2, red) at various shear displacements Δy. **b**, Topological phase diagram showing the WP separation as a function of shear displacement, where the number of WPs changes between zero (0 WPs), four (4 WPs) and eight (8 WPs). **c**, Time trace of WP separation upon launching the shear motion using 23-THz pump pulses at fields of 2.5 MV cm^{-1} (red) and 7.5 MV cm^{-1} (blue). At low fluence, the time trace shows a clear oscillation, increasing and decreasing the WP separation. At high fluence, the WPs mostly annihilate each other. **d**, SHG intensity polar plot of equilibrium WTe_2 (blue), pumped WTe_2 (red) and centrosymmetric $MoTe_2$ (yellow) at various SHG-polarizer angles, where the transmission axis is aligned horizontally at 0° or vertically at 90°. **e**, Pump-induced SHG time traces of WTe_2 at various pump field strengths. Here, the pump pulse has a wavelength of 2.1 μm (polarized at 45° off the horizontal axis), the incident probe pulse has a wavelength of 800 nm and is vertically polarized, and the crystallographic *a* axis is aligned horizontally. The SHG results show that WTe_2 exhibits a transition towards a centrosymmetric, topologically trivial phase.

topologically distinct phases, which are potentially useful for applications. In addition, it can be used to stabilize emergent topological phases at non-equilibrium in otherwise trivial materials, thus diversifying the class of topological materials.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0809-4.

Received: 20 April 2018; Accepted: 25 October 2018;
Published online 2 January 2019.

1. Hasan, M. Z. & Kane, C. L. Colloquium: topological insulators. *Rev. Mod. Phys.* **82**, 3045–3067 (2010).
2. Burkov, A. A. Topological semimetals. *Nat. Mater.* **15**, 1145–1148 (2016).
3. Jia, S., Xu, S. Y. & Hasan, M. Z. Weyl semimetals, Fermi arcs and chiral anomalies. *Nat. Mater.* **15**, 1140–1144 (2016).
4. Roushan, P. et al. Topological surface states protected from backscattering by chiral spin texture. *Nature* **460**, 1106–1109 (2009).
5. Alpichshev, Z. et al. STM imaging of electronic waves on the surface of Bi_2Te_3: topologically protected surface states and hexagonal warping effects. *Phys. Rev. Lett.* **104**, 016401 (2010).
6. Qian, X. et al. Quantum spin Hall effect in two-dimensional transition metal dichalcogenides. *Science* **346**, 1344–1347 (2014).
7. Cortijo, A. et al. Elastic gauge fields in Weyl semimetals. *Phys. Rev. Lett.* **115**, 177202 (2015).

© 2019 Springer Nature Limited. All rights reserved.

8. Grushin, A. G. et al. Inhomogeneous Weyl and Dirac semimetals: transport in axial magnetic fields and Fermi arc surface states from pseudo-Landau levels. *Phys. Rev. X* **6**, 041046 (2016).
9. Pikulin, D. I., Chen, A. & Franz, M. Chiral anomaly from strain-induced gauge fields in Dirac and Weyl semimetals. *Phys. Rev. X* **6**, 041021 (2016).
10. Zeljkovic, I. et al. Strain engineering Dirac surface states in heteroepitaxial topological crystalline insulator thin films. *Nat. Nanotechnol.* **10**, 849–853 (2015).
11. Liu, Y. et al. Tuning Dirac states by strain in the topological insulator Bi_2Se_3. *Nat. Phys.* **10**, 294–299 (2014).
12. Xu, S. Y. et al. Discovery of a Weyl fermion semimetal and topological Fermi arcs. *Science* **349**, 613–617 (2015).
13. Yang, L. X. et al. Weyl semimetal phase in the non-centrosymmetric compound TaAs. *Nat. Phys.* **11**, 728–732 (2015); erratum **11**, 879 (2015).
14. Lv, B. Q. et al. Experimental discovery of Weyl semimetal TaAs. *Phys. Rev. X* **5**, 031013 (2015).
15. Ali, M. N. et al. Large, non-saturating magnetoresistance in WTe_2. *Nature* **514**, 205–208 (2014).
16. Soluyanov, A. A. et al. Type-II Weyl semimetals. *Nature* **527**, 495–498 (2015).
17. Li, P. et al. Evidence for topological type-II Weyl semimetal WTe_2. *Nat. Commun.* **8**, 2150 (2017).
18. Lin, C. L. et al. Visualizing type-II Weyl points in tungsten ditelluride by quasiparticle interference. *ACS Nano* **11**, 11459–11465 (2017).
19. Lv, Y. Y. et al. Experimental observation of anisotropic Adler–Bell–Jackiw anomaly in type-II Weyl semimetal $WTe_{1.98}$ crystals at the quasiclassical regime. *Phys. Rev. Lett.* **118**, 096603 (2017).
20. Yang, J. et al. Elastic and electronic tuning of magnetoresistance in $MoTe_2$. *Sci. Adv.* **3**, eaao4949 (2017).
21. Weathersby, S. P. et al. Mega-electron-volt ultrafast electron diffraction at SLAC National Accelerator Laboratory. *Rev. Sci. Instrum.* **86**, 073702 (2015).
22. Kim, H.-J. et al. Origins of the structural phase transitions in $MoTe_2$ and WTe_2. *Phys. Rev. B* **95**, 180101 (2017).
23. Nicoul, M. et al. Picosecond acoustic response of a laser-heated gold-film studied with time-resolved X-ray diffraction. *Appl. Phys. Lett.* **98**, 191902 (2011).
24. Frigge, T. et al. Optically excited structural transition in atomic wires on surfaces at the quantum limit. *Nature* **544**, 207–211 (2017).
25. Chen, S. Y. et al. Activation of new Raman modes by inversion symmetry breaking in type II Weyl semimetal candidate T′-$MoTe_2$. *Nano Lett.* **16**, 5852–5860 (2016).
26. Zhang, K. et al. Raman signatures of inversion symmetry breaking and structural phase transition in type-II Weyl semimetal $MoTe_2$. *Nat. Commun.* **7**, 13552 (2016).
27. Morimoto, T. & Nagaosa, N. Topological nature of nonlinear optical effects in solids. *Sci. Adv.* **2**, e1501524 (2016).
28. Wu, L. et al. Giant anisotropic nonlinear optical response in transition metal monopnictide Weyl semimetals. *Nat. Phys.* **13**, 350–355 (2017).
29. Levy, N. et al. Strain-induced pseudo-magnetic fields greater than 300 tesla in graphene nanobubbles. *Science* **329**, 544–547 (2010).
30. Gomes, K. K. et al. Designer Dirac fermions and topological phases in molecular graphene. *Nature* **483**, 306–310 (2012).

Acknowledgements This work is supported primarily by the US Department of Energy (DOE), Office of Basic Energy Sciences, Division of Materials Sciences and Engineering, under contract number DE-AC02-76SF00515, the Stanford Linear Accelerator (SLAC) National Accelerator Laboratory, the Stanford Institute for Materials and Energy Sciences (E.J.S., C.M.N., C.D.P., E.M., T.P.D., T.F.H., A.M.L.). E.J.S. acknowledges additional support from Stanford GLAM Postdoctoral Fellowship Program. C.M.N. acknowledges additional support from the National Science Foundation (NSF) through a Graduate Research Fellowship (DGE-114747). T.F.H. acknowledges additional funding for analysis from the Gordon and Betty Moore Foundation EPiQS Initiative through grant number GBMF4545. S.J.P. is supported by the US Department of Energy (DE-SC0012375). M.C.H. is supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences, under award number 2015-SLAC-100238-Funding. B.K.O.-O. acknowledges support from the DOE Office of Science, Fusion Energy Science, under grant number FWP 100182. N.F. acknowledges support from the Stewardship Science Graduate Fellowship programme, provided under cooperative agreement number DE-NA0002135. Synthesis of $MoTe_2$ and sample preparation were supported by the US Department of Energy, DE-SC0016703 (D.R., D.C., A.A., J.H.). L.B. acknowledges the US Army Research Office MURI grant W911NF-11-1-0362. The National High Magnetic Field Laboratory is supported by the NSF through NSF/DMR-1157490, NSF/DMR-1644779 and the State of Florida. First-principles calculations by C.D.P. were supported by the TIMES programme at SLAC. Numerical simulations were performed using computational resources at the National Energy Research Scientific Computing Center (NSERC). The UED work was performed at SLAC MeV-UED, which is supported in part by the DOE BES SUF Division Accelerator & Detector R&D programme, the LCLS Facility and SLAC under contracts DE-AC02-05-CH11231 and DE-AC02-76SF00515. The authors thank D. Pikulin and B. Moritz for discussions and G. Stewart for the illustration of the UED setup.

Reviewer information *Nature* thanks C. Ropers and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions E.J.S., C.M.N. and A.M.L. conceived the experiments, analysed and interpreted the data; E.J.S. and C.M.N. performed the SHG experiments, analysed and interpreted the data with A.M.L. and T.F.H.; E.J.S., C.M.N., A.M.L. and C.D.P. wrote the manuscript with input from all authors; E.J.S. and C.M.N. performed the UED experiments with S.J.P., X.S., J.Y., R.L., S.W., E.M. and X.J.W.; B.K.O.-O., M.C.H. and A.H.R. implemented the terahertz setup; D.R., L.B., C.M.N., E.J.S., N.F., D.C., A.A., J.H. and T.F.H. synthesized the crystals and prepared the samples; C.D.P. and T.P.D. performed the first-principles calculations; X.J.W. led the SLAC 3-MeV UED and terahertz-pump UED-probe development. All authors discussed the results and contributed to the writing of the manuscript.

Competing interests E.J.S., C.M.N., C.D.P., X.J.W. and A.M.L. have submitted a patent application ("Fast topological switch using strained Weyl semimetals"; US number 62/726,893) that covers a specific aspect of the manuscript. The other authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0809-4.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0809-4.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.M.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

Sample synthesis and preparation. High-quality single crystals of WTe_2 were synthesized using a self-flux method in excess of Te. W 99.999% and Te 99.9999% powders were placed in a quartz ampoule in a ratio of 1:25, heated to 1,100 °C and held at this temperature for three days. Subsequently, the ampoule was slowly cooled to 525 °C over two weeks and centrifuged. The 'as-harvested' single crystals were then annealed for two days at a temperature of 425 °C under a temperature gradient to remove excess Te. To prepare the synthesized samples for UED experiments, WTe_2 was mechanically exfoliated onto a SiO_2/Si substrate using a standard mechanical exfoliation technique. From the exfoliated crystal, the samples were selected for subsequent transfer by their size (>50 μm in the lateral dimension) and thickness (>50 nm). After verifying the thickness using an atomic force microscope, poly(propylene carbonate) (PPC) in anisole solution (15% PPC by weight) was spun onto the WTe_2-covered SiO_2/Si substrate at a rate of 1,500 r.p.m. with an acceleration of 1,000 r.p.m. s^{-1} for 1 min and then heated to 80 °C for 2 min on a hotplate. The PPC film and the WTe_2 crystal were then peeled from the substrate and suspended over a hotplate with the WTe_2 facing up. A 50-nm-thick Si_3N_4 transmission-electron-microscopy membrane was then aligned over the suspended crystal using an optical microscope, and placed on the PPC film while raising the temperature to 115 °C to induce contact between the WTe_2 crystal and the membrane. The sample was then soaked in acetone for 10 min to remove the PPC, gently rinsed with isopropyl alcohol and dried under a flow of nitrogen gas, thus completing the transfer.

SLAC 3-MeV UED setup. We used a relativistic UED technique to reconstruct the shear motion and crystallographically measure the corresponding atomic displacements through the measurement of more than 200 Bragg peaks (Fig. 1b). The electron beam is generated using a frequency-tripled Ti:sapphire laser by excitation of a copper photocathode and rapidly accelerated to 3 MeV in radiofrequency electric fields[21]. The pulse duration of the electron beam at the sample position is 150 fs. Magnetic lenses are used to steer and focus the electron beam[31] onto the sample, an exfoliated single-domain crystal, with a spot size of 100 μm. The diffracted electron beam is captured in a transmission geometry using an electron-multiplying CCD camera. We used two different pump excitation schemes in this experiment: a quasi-single-cycle excitation at 1–5 THz and a few-cycle excitation at 23 THz. The arrival time of the electron beam (probe) could be adjusted with respect to the terahertz pulses (pump) using an optical delay stage.

Ultrafast terahertz sources. Quasi-single-cycle terahertz pulses were generated by optical rectification of 1,350-nm near-infrared laser pulses in the organic nonlinear crystals DSTMS (4-N,N-dimethylamino-4′-N′-methyl-stilbazolium 2,4,6-trimethylbenzenesulfonate)[32] and OH-1 (2-(3-(4-hydroxystyryl)-5,5-dimethylcyclohex-2-enylidene)malononitrile)[33]. The 1,350-nm near-infrared pulses were generated from an 800-nm Ti:sapphire laser system in a three-stage optical parametric amplifier system (Light Conversion HE-TOPAS) and had pulse energies up to 2 mJ and a pulse duration of about 50 fs.

The terahertz field was brought to an intermediate focus with an off-axis parabolic mirror with 2-inch (50.8 mm) focal length, collimated with a second mirror with 6-inch (152.4 mm) focal length. The collimated beam was transported into the UED diffraction chamber via a polymer window and focused with an off-axis parabolic mirror with 3-inch (76.2 mm) focal length inside the chamber. The terahertz field was characterized at the sample location by electro-optical sampling using a split-off portion of the 800-nm laser and a 50-μm-thick 110-cut GaP crystal. The observed peak field strength of the DSTMS-generated terahertz pulse was 650 kV cm^{-1}, with the spectrum centred at about 3 THz when using an 8-mm-diameter 450-μm-thick crystal and input pump pulse energy of about 1 mJ. With a 10-mm-diameter clear aperture and a 500-μm-thick OH-1 crystal, the peak field strength was 500 kV cm^{-1}, with the spectrum peaked at 1.5 THz and sizeable spectral components extending to about 3.5 THz (Extended Data Fig. 2). The polarization of the terahertz pulses was linear and the polarization angle could be changed arbitrarily by rotating the generation crystal together with the pump-pulse polarization. A detailed discussion of the experimental apparatus can be found in ref. [34].

Mid-infrared (MIR) pulses with 13-μm wavelength (23 THz frequency) were generated by difference-frequency generation in GaSe from the signal and idler of the same optical parametric amplifier system (Light Conversion HE-TOPAS), driven by 800-nm pulses with duration of about 130 fs. Here the signal and idler wavelengths were 1,505 nm and 1,705 nm, respectively. The MIR beam was transported into the experimental chamber through a 3-mm-thick KRS-5 window and focused with an off-axis parabolic mirror with 3-inch (76.2 mm) focal-length. A pair of holographic wire-grid polarizers (Thorlabs WP25H-K) was used to attenuate the pulse energy to the desired level. The pulse duration of the MIR pulses was of the order of 300 fs after taking into account dispersion. MIR spot-size measurements at the sample position were obtained with a DataRay WinCamD beam profiler.

Structure factor calculation with interlayer shear displacement. The intensity of a Bragg peak, $I \propto |S|^2$, can be calculated using the general form of the structure factor

$$S(hkl) = \sum_j f_j \exp[-i2\pi(hx_j + ky_j + lz_j)] \tag{2}$$

where the summation runs over all atoms in the unit cell (four W and eight Te atoms), f_j is the atomic scattering factor for the jth atom, $\boldsymbol{r}_j = x_j\hat{\boldsymbol{a}} + y_j\hat{\boldsymbol{b}} + z_j\hat{\boldsymbol{c}}$ is the vector position of the atom in the unit cell ($0 \le x, y, z \le 1$) and (hkl) are the usual Miller indices. Because we use transmission geometry at the [001] zone axis, the diffraction image shows only the $l = 0$ peaks, that is, (hk0). We calculate the peak intensity modulation $\Delta I/I_0$ by introducing the top-layer shear displacement Δy with respect to the bottom layer into the structure factor

$$\begin{aligned} S(\Delta y) = \exp(-i2\pi k\Delta y) \sum_{\text{top}} f_j \exp[-i2\pi(hx_j + ky_j)] \\ + \sum_{\text{bottom}} f_j \exp[-i2\pi(hx_j + ky_j)] \end{aligned} \tag{3}$$

We can obtain a more symmetric expression by having the shear displacement shared equally between the two layers ($\Delta y/2$), that is, by multiplying with a common phase factor $\exp(+i\pi k\Delta y)$ and by using the underlying crystal symmetry through (i) a reflection with respect to the b–c mirror plane ($x \to -x$) and (ii) a non-symmorphic C_2 transformation. The latter consists of a reflection with respect to the a–c mirror plane ($y \to -y$) and a translation along the a–c axis (+0.5, 0, +0.5). These symmetry operations project each atom from the bottom layer to the top layer

$$\begin{aligned} S(\Delta y) &= \exp(-i\pi k\Delta y) \sum_{\text{top}} f_j \exp[-i2\pi(hx_j + ky_j)] \\ &\quad + \exp(+i\pi k\Delta y) \sum_{\text{top}} f_j \exp[+i2\pi(hx_j + ky_j)] \\ &= 2\sum_{\text{top}} f_j \cos[2\pi(hx_j + ky_j) + \pi k\Delta y] \end{aligned} \tag{4}$$

Now the summation runs over all atoms in the top half of the unit cell (two W and four Te atoms). To compare this with our experiment, we calculated the change of peak intensity $\Delta|S(\Delta y)|^2 = |S(\Delta y)|^2 - |S(0)|^2$ using known x_j and y_j values[16], as shown in Fig. 2b. The structure factors f_j were calculated using X-ray-scattering factors from published analytical fits[35,36] converted to electron-scattering factors using the Mott–Bethe formula[37]. Here, we define a positive shear displacement ($\Delta y > 0$) as shown in the inset of Fig. 2e.

Fitting of the structural-factor modulations. For each time point Δt, the mean-square error (P) is calculated for a range of shear displacements Δy for a selection of m Bragg peaks (hkl). In addition, an anisotropic (elliptical) Debye–Waller factor is included to account for heating effects in the sample, which is considerably smaller than those from the shear displacements due to the low pump photon energy (terahertz). The mean squared error is

$$P(\Delta y, \langle u_a^2\rangle, \langle u_b^2\rangle) = \frac{1}{m}\sum_{hkl}\left[\frac{\Delta I}{I_0}(hkl, \Delta y, \langle u_a^2\rangle, \langle u_b^2\rangle)_{\text{sim}} - \frac{\Delta I}{I_0}(hkl)_{\text{exp}}\right]^2$$

where $\langle u_a^2\rangle$ and $\langle u_b^2\rangle$ are the mean-square atomic displacements along the a and b axes, respectively, which affect the intensity of a Bragg peak by the Debye–Waller relation, $I = I_0 \exp\left[-\frac{1}{2}(Q_a^2\langle u_a^2\rangle + Q_b^2\langle u_b^2\rangle)\right]$, with a time constant determined by the (400) Bragg peak. Here, Q_a and Q_b are the projections of $\boldsymbol{Q}$, the reciprocal lattice vector of the Bragg peak, along the a and b axes. The simulated intensity change, $(\Delta I/I_0)_{\text{sim}}$, has the form

$$\frac{\Delta I}{I_0}(hkl, \Delta y, \langle u_a^2\rangle, \langle u_b^2\rangle)_{\text{sim}} = \frac{|S(\Delta y)|^2\, e^{-\frac{1}{2}(Q_a^2\langle u_a^2\rangle + Q_b^2\langle u_b^2\rangle)} - |S(0)|^2}{|S(0)|^2}$$

Here, $S(\Delta y)$ is the structure factor calculated for a given Δy, as discussed in the main text, and $S(0)$ is the structure factor calculated for the undistorted structure. The values for each parameter Δy, $\langle u_a^2\rangle$ and $\langle u_b^2\rangle$ at every Δt are optimized by minimizing P. In other words, the fitting procedure is performed to minimize the peak intensity difference between experiment and simulation and is averaged across many Bragg peaks.

Estimating the effective terahertz-induced hole doping. We use a simple Drude model to estimate the effective hole doping. The fraction of electrons that contribute to the resulting current is v/v_F, where v is the drift velocity and v_F is the Fermi velocity. The drift velocity can be estimated through $v = eE\tau/m$, where e is the electron charge, E is the applied electric field, τ is the scattering time and m is the effective mass. By using the reported values[17] of $m \approx 0.4m_e$ (m_e, electron mass),

© 2019 Springer Nature Limited. All rights reserved.

$v_F \approx 3 \times 10^5$ m s^{-1}, a typical value of $\tau \approx 10$ fs in a semimetal and $E \approx 1$ MV cm^{-1}, we found that $v \approx v_F$. For a hole pocket with carrier density of $n_0 \approx 7 \times 10^{19}$ cm^{-3} in WTe_2[38], this is comparable with an effective hole doping density required for the Td–1T′ phase transition[22] as the impulsive driving force for the interlayer shear motion. Moreover, recent experiments reported considerably larger scattering time values of $\tau > 100$ fs in WTe_2[17,39]. This suggests that an even larger electron density can be transiently transferred away from the topmost valence band by the terahertz pump field to induce the interlayer shear motion. We note that the terahertz-induced effective doping serves as an impulsive driving force to kick-start the shear mode and does not require the excited carriers to be maintained during the long-lived metastable phase. Such a metastable phase persists for a time that is determined by the energy barrier of the local potential minimum and thermal fluctuations.

DFT analysis. DFT[40,41] simulations of WTe_2 were carried out to ascertain the energetics of the experimentally observed 0.24-THz shear mode of interest (Fig. 2e) and the Brillouin-zone (BZ) motion of WPs resulting from atomic displacements associated with this mode. The topological characteristics of the electronic band structure of Td WTe_2 have been explored previously[16,22]. In particular, Soluyanov et al.[16] investigated the type II Weyl semimetal character of Td WTe_2 partly on the basis of DFT band-structure simulations carried out at the experimentally observed geometry[42]. More recently, Kim et al.[22] studied the effect of geometry optimization, within the van der Waals (vdW) DFT framework[43], on the WPs in Td WTe_2. They found that the stability of Weyl nodes in WTe_2 is sensitive to lattice parameter differences of the order of 1%; in fact, owing to slight (about 1%) inaccuracies in the vdW–DFT-predicted a and c parameters, WPs of opposite chiralities annihilate at the vdW–DFT equilibrium geometry of WTe_2, rendering it a trivial semimetal. To recover distinct Weyl nodes in the BZ, small strains need to be applied[22].

To circumvent the issues related to the sensitive dependence on specific vdW–DFT geometries, we initially employ as reference the experimental geometry[42] used in the work of Soluyanov et al.[16] to carry out our analysis of the BZ motion of the Weyl nodes in WTe_2 under the influence of shear-mode displacements along the y axis (crystallographic b axis). This geometry is characterized by an orthorhombic (Td) lattice with parameters $a = 3.477$ Å, $b = 6.249$ Å and $c = 14.018$ Å. W and Te atoms occur at the 2a Wyckoff positions parameterized as $(0, y, z)$ and $(1/2, -y, z+1/2)$ for values of y, z given in Extended Data Table 1a (reproduced from Soluyanov et al.[16]). Simulations in this context are carried out using the DFT framework as implemented in the Vienna ab initio simulation package[44] (VASP) version 5.4.1. All calculations include spin–orbit coupling within the non-collinear DFT formalism. Exchange-correlation effects are treated at the level of the generalized gradient approximation through the PBE[45] functional. Projector-augmented-wave[46] potentials with valence electronic configurations of $\{6s^2, 5d^4\}$ for W and $\{5s^2, 5p^4\}$ for Te are employed in conjunction with a plane-wave energy cutoff parameter of 260 eV. For electron density convergence, a Γ-centred $12 \times 10 \times 6$ k-point grid and Gaussian smearing with a smearing parameter of 0.05 eV are used. The simulation parameters here are similar to those adopted in Soluyanov et al.[16]. WP positions in the $k_z = 0$ plane of the BZ of WTe_2 are subsequently identified through band-structure calculations employing a dense $43 \times 85 \times 1$ k-point mesh spanning a sub-region of the $(k_x, k_y, k_z = 0)$ plane, as shown in Fig. 4a. At the employed geometry, we identify two WPs (WP1, WP2) in the first quadrant of the $k_z = 0$ plane of the BZ at the coordinates shown in Extended Data Table 1b. The remaining six points in the other three quadrants are related to these by reflections in the $k_x = 0$ and $k_y = 0$ planes[16]. The coordinates identified here are similar to those reported in the literature[16,22].

Phonon band-structures are calculated within the frozen-phonon finite-difference approach using the *phonopy*[47] code interfaced with VASP. In Extended Data Fig. 3a, the phonon band dispersions of Td WTe_2 at the experimental reference geometry are plotted along high-symmetry lines in the $k_z = 0$ plane. At the Γ point, the low-energy mode at 0.29 THz is unambiguously identified as the relevant interlayer shear mode of interest, which corresponds to the 0.24-THz mode observed experimentally. This identification is facilitated by the absence of any other modes at similar energies and also by inspecting the atomic position modulations associated with the mode eigenvector. As denoted in Fig. 2e of the main text, this mode predominantly involves a relative shear displacement along the b axis of adjacent WTe_2 layers in the unit cell. To investigate the motion of WPs in the BZ induced by this mode, first a number of unit-cell geometries modulated along the mode displacement coordinate are generated using the *phonopy* code. Subsequently, the DFT electronic-band structures corresponding to these modulated geometries are calculated, and the positions of WPs in the BZ are mapped as a function of displacement along the mode coordinate, as shown in Fig. 4a. As explained in the main text in connection with Fig. 4a, depending on the sign of the shear-mode displacement along the b axis, pairs of WPs either move closer or farther apart in the BZ, leading to different kinds of topological transitions.

In topological semimetals such as WTe_2, the positions of WPs in the BZ can be tuned by applying strain. Soluyanov et al.[16] explored changes in the relative positions of Weyl nodes in WTe_2 induced by uniaxial tensile and compressive strains applied along different crystallographic axes. They report that whereas stretching along the a axis leads to annihilation of all pairs of Weyl nodes, compressive strain along this direction leads to increased separation within each pair of WPs, until half the points eventually annihilate on the $k_y = 0$ mirror plane, leading to a state where only four Weyl nodes survive. As explained in the analysis accompanying Fig. 4, we find that a similar motion of the WPs can be induced via an alternative mechanism, namely, terahertz-pump-induced phonon modulations associated with the 0.24-THz shear mode along the b axis. It is therefore instructive to evaluate the relative energy cost associated with these two approaches to tuning the topological properties of WTe_2. To this end, we compare DFT total energies of different strained and phonon-modulated structures associated with the WP motion described in Fig. 4. Calculating the total energy cost under different deformations of the lattice requires identifying the minimum energy point associated with the equilibrium geometry. Therefore, for this analysis, we first carry out geometry optimizations within the dispersion corrected DFT-D3[48] framework including spin–orbit coupling, as implemented in VASP. A plane-wave cutoff of 400 eV, in conjunction with a Γ-centred $16 \times 9 \times 4$ k-point grid, is employed in this instance. As before, exchange-correlation effects are modelled using the PBE functional; additionally, dispersion corrections are incorporated at the level of the DFT-D3[48] approximation. In particular, for reasons explained below, two different forms for the dispersion corrections are investigated, namely, DFT-D3[11] (labelled D3) and DFT-D3 with Becke–Johnson damping[49] (labelled D3-BJ). The lattice parameters and relevant interlayer distances d_1, d_2, d_3 (see Fig. 2e) in Td WTe_2 predicted by the two methods are listed in Extended Data Table 1c and compared to experimental values.

The last column of Extended Data Table 1c shows the calculated and experimental phonon frequency for the b-axis shear mode. As mentioned earlier, this mode is well approximated as a rigid relative motion along the b axis of the two WTe_2 layers in the unit cell (see Extended Data Fig. 3a). The potential energy associated with this motion is predominantly determined by the interlayer vdW interaction, at least for small displacements, and therefore exhibits a strong dependence on the way that the dispersion corrections are approximated in DFT. We note that whereas the D3 method yields lattice parameters that are within 0.5% of the experimental values, the b-axis shear-mode phonon frequency is underestimated by about 37%. On the other hand, the D3-BJ approximation predicts a mode frequency that is overestimated by about 40%. In the harmonic approximation within which the phonon frequency is estimated, this corresponds to a picture where the potential energy surface associated with displacement along the shear mode is too shallow and too steep in the D3 and D3-BJ approximations, respectively. Although neither method is quantitative, we expect that the correct description lies between the two limits represented by D3 and D3-BJ. With this understanding, we provide here an order-of-magnitude comparison between the relevant strain and phonon modulation energies for driving WP motion in WTe_2. With reference to the equilibrium geometry, the total energy cost per unit cell as a function of applied compressive strain along the a axis is shown in Extended Data Fig. 3b. Similarly, the total energy cost associated with structural modulation by the b-axis shear phonon as a function of displacement along the mode coordinate is shown in Extended Data Fig. 3c. The displacement (in picometres) along the b axis of one of the W atoms is used as a proxy for the mode coordinate. To annihilate two pairs of Weyl nodes at the $k_y = 0$ mirror plane, either a 2% a-axis compressive strain[16] or a negative displacement of about 12 pm in the sense of Fig. 2e is required. On a per-unit-cell basis, in the D3 (D3-BJ) approximation the former mechanism has an energy cost of 32 meV (39 meV) whereas the latter costs 0.33 meV (3.6 meV). Thus, for driving the topological transition from eight to four Weyl nodes, the lattice strain mechanism is 1–2 orders of magnitude more expensive compared to the phonon-driven mechanism in terms of energy. This is expected because an a-axis strain involves compressing or elongating strong covalent bonds within each layer of WTe_2, whereas b-axis shear mode involves primarily interlayer interactions, which are weaker.

Longitudinal acoustic-wave timescale. In WTe_2, the stable Td phase appears as an orthorhombic unit cell in which the b and c axes form an angle of $\theta = 90°$, whereas the 1T′ phase appears as a monoclinic unit cell with a 94° angle (Extended Data Fig. 1). Hence, we can expect the Td–1T′ transition in WTe_2 to be limited by the time required for the development of an overall shear across the sample thickness (c axis), as determined by the transverse acoustic group velocity. Although we are not aware of any prior measurements on the transverse acoustic velocity, we can take the longitudinal acoustic speed of sound[50] (2,000 m s^{-1}) as a first approximation. Hence, we can estimate the build-up time using $t = d/v$, where $d = 50$ nm is the sample thickness and $v = 2{,}000$ m s^{-1} is the speed of sound, and obtain $t = 25$ ps. This is in good agreement with our observation (Fig. 2f, blue line). We can take a separate confirmation of the longitudinal acoustic speed of sound by tilting the sample (pitch of −8.3°, yaw of −13.9°), providing sensitivity to the acoustic breathing mode oscillations (period of $2d/v$) across the sample thickness. The structural-factor modulation $\Delta I/I_0$ of peak $(13\bar{3})$ shows oscillations with period of 38 ps (Extended Data Fig. 4). Given the period of this oscillation,

© 2019 Springer Nature Limited. All rights reserved.

we compute the speed of sound to be $v \approx 2{,}600$ m s^{-1}, in reasonable agreement with the above estimate.

Terahertz-induced shear motion in $MoTe_2$. We discussed in the main text how the shear motion in WTe_2 is driven through a transient hole doping induced by the terahertz field. This interpretation is motivated by a theoretical prediction that upon hole doping the Td phase becomes unstable against the 1T′ phase[22]. A particularly strong indication from our experiment is that at any terahertz-field polarization, the initial shear motion always occurs along the pathway towards a phase transition from the Td phase to the 1T′ phase. Hence, this process must be very sensitive to the initial structural phase of the sample, that is, it should only occur in the Td phase and not in the 1T′ phase—a feature that can be tested. In particular, we should expect the absence of terahertz-induced shear motion in 1T′ $MoTe_2$. $MoTe_2$ and WTe_2 have similar structural and electronic properties; unlike WTe_2, however, $MoTe_2$ can appear in two structural phases: the Td phase below 200 K, the 1T′ phase above 250 K and a mixed Td–1T′ phase[25] at 200–250 K. We performed similar terahertz-pump UED-probe experiments on $MoTe_2$ at sample temperatures of 28 K (Td) and 300 K (1T′) (Extended Data Fig. 5). We found that the interlayer shear oscillations only occur in the Td phase (0.37 THz), and not in the 1T′ phase of $MoTe_2$, where only a small heating (Debye–Waller) effect is observed. This is consistent with the picture that terahertz-induced hole doping stabilizes the 1T′ phase over the Td phase. That is, if we start with the Td phase, the relative energy order is reversed upon terahertz-induced hole doping and shear motion is launched; but if we start with the 1T′ phase, the relative energy order does not change and no shear motion is observed.

Transition region at intermediate pump fluence. The UED time traces (Fig. 2f) show that the oscillation period is slightly longer at larger pump field strengths. To investigate this transition region, we carried out an additional terahertz-pump UED-probe experiment at smaller pump fluence intervals (Extended Data Fig. 6a, b). As we can see, the oscillation period increases at higher pump fluences. This observation indicates a nonlinear phonon softening towards a new metastable centrosymmetric structure.

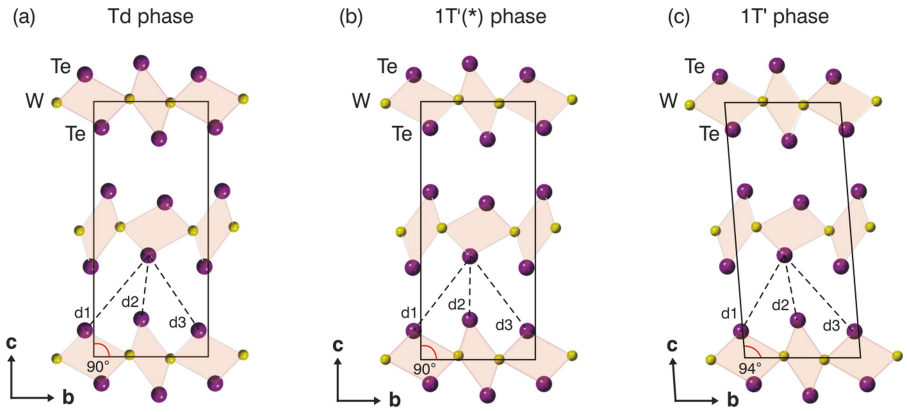
Lifetime of excited electrons. The lifetime of excited electrons can be determined from the pump-induced probe reflectivity as a function of time. We carried out an optical pump-probe experiment using 2.1-μm pump and 800-nm probe pulses on a WTe_2 sample (Extended Data Fig. 7a). We note that we replace the role of the terahertz pump with a 2.1-μm pump because the terahertz pump setup that we used for UED requires a special laser specification and is currently not accessible for optical reflectivity experiments. Nevertheless, measurements with 2.1-μm pump pulses can also induce the shear oscillations that we obtained using terahertz pump pulses (Extended Data Fig. 7b). Extended Data Fig. 7a shows an abrupt pump-induced change of probe reflectivity within the first 5 ps and a stable finite reflectivity afterwards for a timescale longer than 50 ps. The first 5 ps can be attributed to relaxation of hot carriers towards a new quasi-equilibrium state. Afterwards, the carriers remain in the new equilibrium state for longer than 50 ps.

Data availability

The data that support the findings of this study are available from the corresponding author on reasonable request.

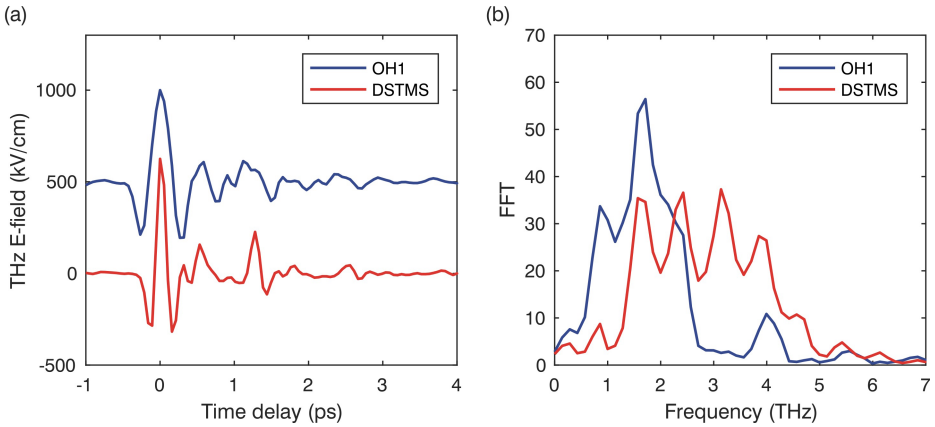
31. Shen, X. et al. Femtosecond mega-electron-volt electron microdiffraction. *Ultramicroscopy* **184**, 172–176 (2018).
32. Yang, Z. et al. Large-size bulk and thin-film stilbazolium-salt single crystals for nonlinear optics and THz generation. *Adv. Funct. Mater.* **17**, 2018–2023 (2007).
33. Brunner, F. D. et al. A hydrogen-bonded organic nonlinear optical crystal for high-efficiency terahertz generation and detection. *Opt. Express* **16**, 16496–16508 (2008).
34. Ofori-Okai, B. K. et al. A terahertz pump mega-electron-volt ultrafast electron diffraction probe apparatus at the SLAC Accelerator Structure Test Area facility. *J. Instrum.* **13**, P06014 (2018).
35. Waasmaier, D. & Kirfel, A. New analytical scattering-factor functions for free atoms and ions. *Acta Crystallogr. A* **51**, 416–431 (1995).
36. Su, Z. & Coppens, P. Relativistic X-ray elastic scattering factors for neutral atoms $Z = 1$–54 from multiconfiguration Dirac–Fock wavefunctions in the 0–12 Å$^{-1}\sin\theta/\lambda$ range, and six-Gaussian analytical expressions in the 0–6 Å$^{-1\ \text{range}}$. *Acta Crystallogr. A* **53**, 749–762 (1997).
37. De Graef, M. *Introduction to Conventional Transmission Electron Microscopy* (Cambridge Univ. Press, Cambridge, 2003).
38. Zhu, Z. et al. Quantum oscillations, thermoelectric coefficients, and the Fermi surface of semimetallic WTe_2. *Phys. Rev. Lett.* **114**, 176601 (2015).
39. Homes, C. C., Ali, M. N. & Cava, R. J. Optical properties of the perfectly compensated semimetal WTe_2. *Phys. Rev. B* **92**, 161109 (2015).
40. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Phys. Rev.* **136**, B864–B871 (1964).
41. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133–A1138 (1965).
42. Mar, A., Jobic, S. & Ibers, J. A. Metal-metal vs tellurium-tellurium bonding in WTe_2 and its ternary variants $TaIrTe_4$ and $NbIrTe_4$. *J. Am. Chem. Soc.* **114**, 8963–8971 (1992).
43. Hamada, I. van der Waals density functional made accurate. *Phys. Rev. B* **89**, 121103 (2014).
44. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
45. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
46. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
47. Togo, A. & Tanaka, I. First principles phonon calculations in materials science. *Scr. Mater.* **108**, 1–5 (2015).
48. Grimme, S. et al. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. *J. Chem. Phys.* **132**, 154104 (2010).
49. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).
50. Zhou, Y. et al. Direct synthesis of large-scale WTe_2 thin films with low thermal conductivity. *Adv. Funct. Mater.* **27**, 1605928 (2017).

© 2019 Springer Nature Limited. All rights reserved.



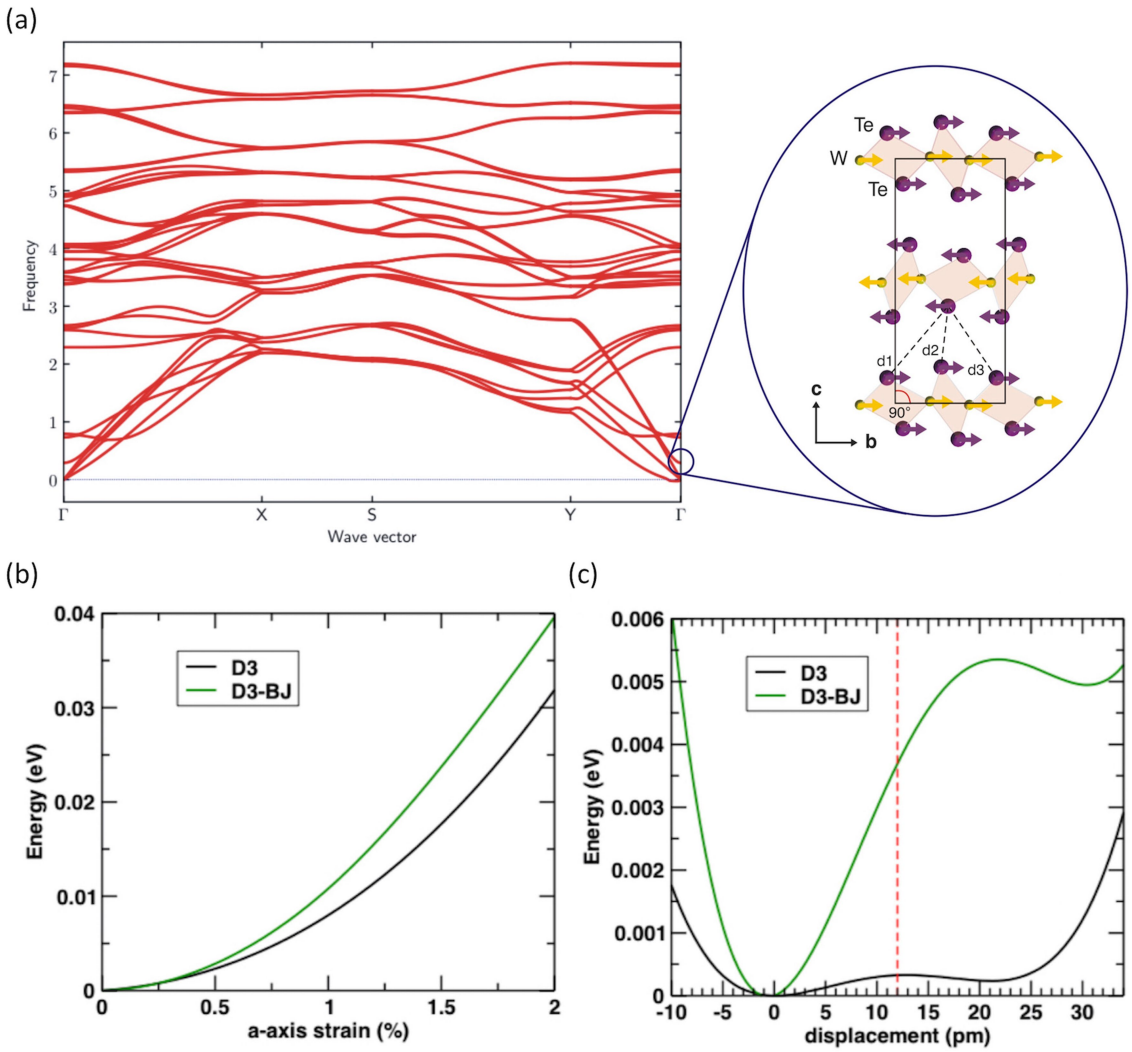
Extended Data Fig. 1 | Lattice structure variants of WTe_2. a, The Td phase has an orthorhombic non-centrosymmetric unit cell with b–c angle of 90° and bond length $d_1 > d_3$. **b**, The 1T′(*) phase has an orthorhombic centrosymmetric unit cell with a b–c angle of 90° and bond length $d_1 = d_3$. **c**, The 1T′ phase has a monoclinic centrosymmetric unit cell with a b–c angle of about 94° and bond length $d_1 < d_3$.

© 2019 Springer Nature Limited. All rights reserved.



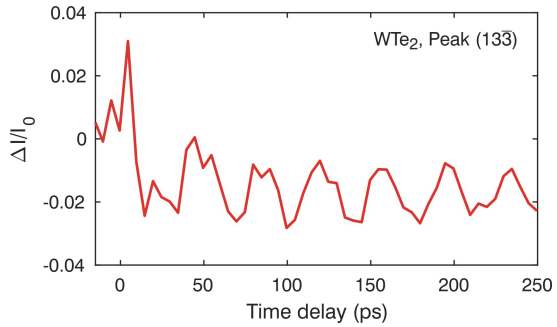
Extended Data Fig. 2 | Electro-optical sampling data of the terahertz pump pulses. a, Time trace of terahertz electric field, generated using OH-1 and DSTMS crystals. Curves in **a** are offset for clarity. **b**, Frequency bandwidth of the terahertz field, calculated using the Fourier transform of **a**.

© 2019 Springer Nature Limited. All rights reserved.



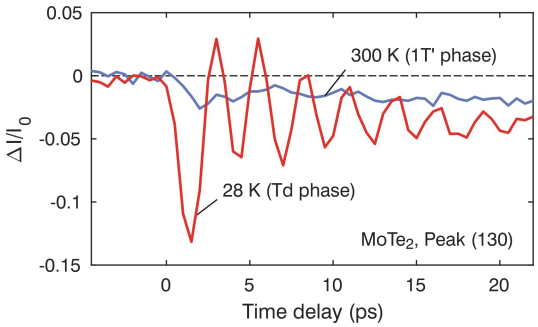
Extended Data Fig. 3 | Calculated phonon dispersion of Td WTe_2 and energy potential as a function of lattice deformation. **a**, Dispersions for wave vectors along high-symmetry lines in the $k_z = 0$ plane are shown. The schematic on the right shows the interlayer shear motion as rigid displacements between alternating WTe_2 layers. **b**, Energy as a function of uniaxial strain applied along the *a* axis. We used two different forms for the dispersion corrections, namely, DFT-D3 (labelled D3) and DFT D3 with Becke–Johnson damping (labelled D3-BJ). These two corrections result in slightly different lattice constants, as shown in Extended Data Table 1c, and yield potential energy surfaces that are too shallow and too steep in the D3 and D3-BJ approximations, respectively. The correct description lies between the two limits represented by D3 and D3-BJ. **c**, Energy as a function of displacement along the shear-mode coordinate. The red dashed line indicates the displacement at which two pairs of Weyl nodes annihilate at the $k_y = 0$ mirror plane (see Fig. 4).

© 2019 Springer Nature Limited. All rights reserved.



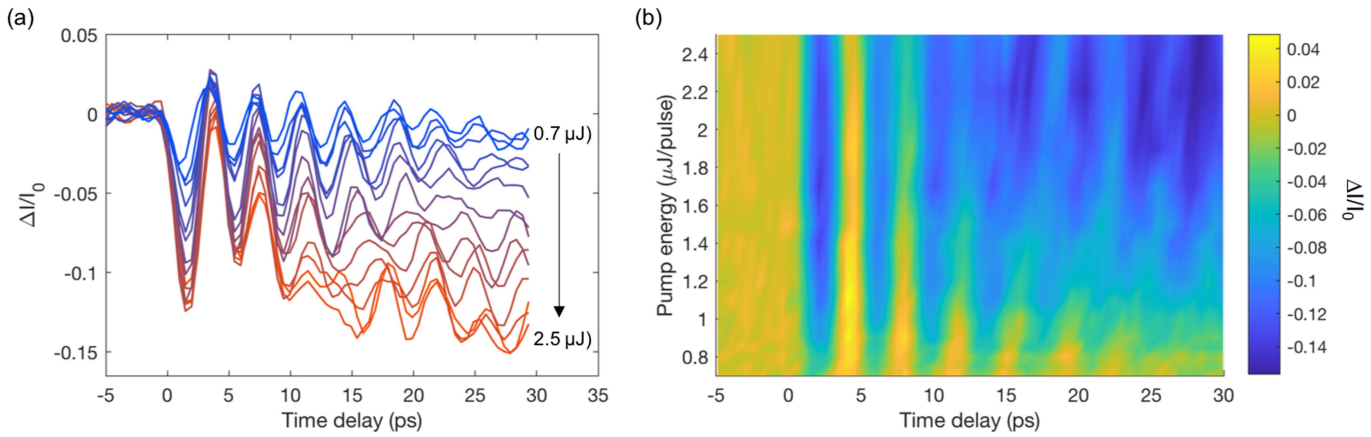
Extended Data Fig. 4 | Transverse acoustic propagation dynamics in Td-WTe_2. The structure-factor modulation is monitored using a terahertz-pump UED probe.

© 2019 Springer Nature Limited. All rights reserved.



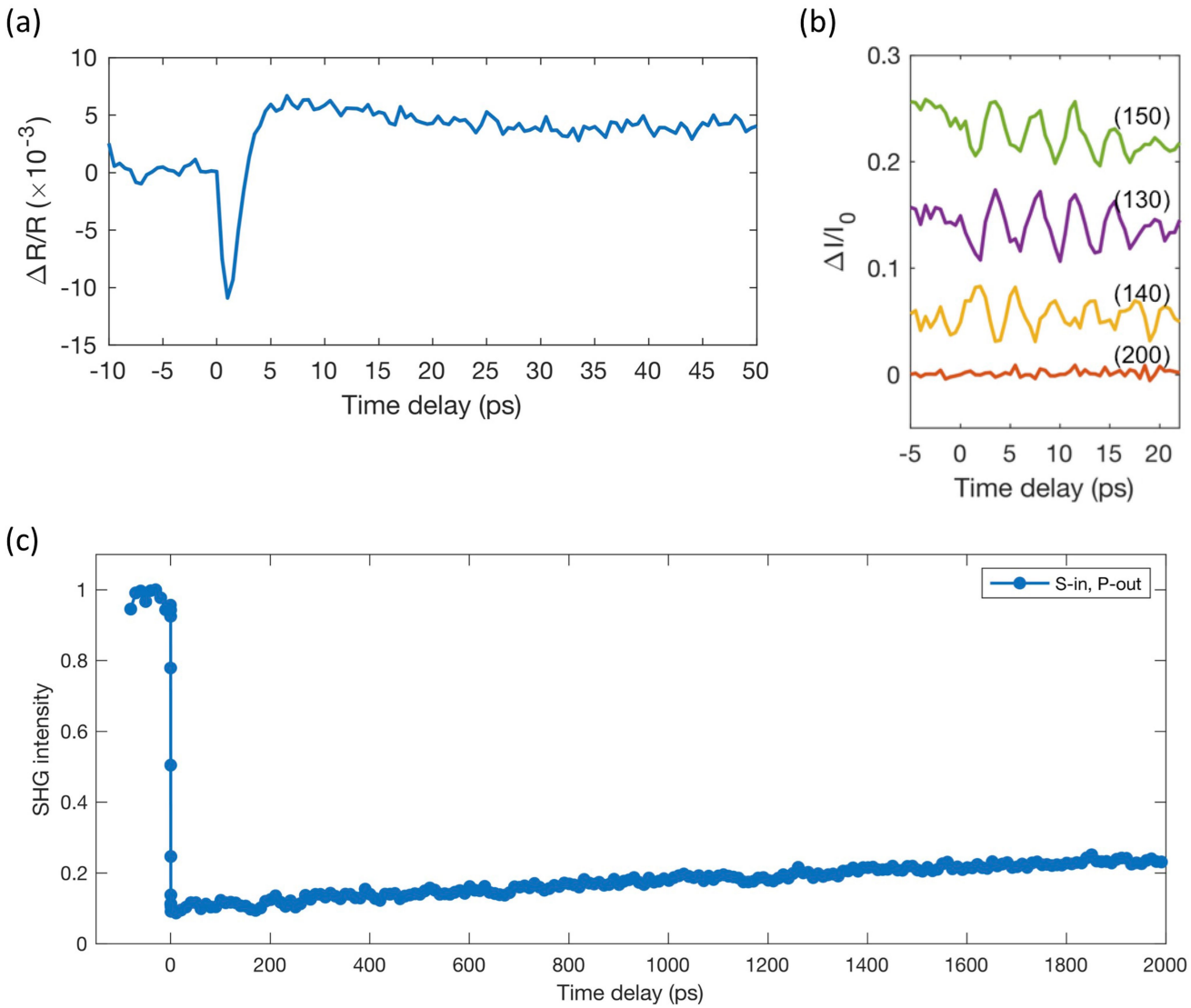
Extended Data Fig. 5 | The emergence of terahertz-induced shear oscillations in Td $MoTe_2$, but not in 1T′ $MoTe_2$. The structure-factor modulations are monitored using a terahertz-pump UED probe.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Additional terahertz-pump UED-probe measurements at increasing pump fluence. a, Intensity changes of the (130) Bragg peak show the interlayer shear oscillation, which exhibits a phonon softening at larger pump fluences. This demonstrates the evolution towards switching behaviour in the transition region. **b**, Surface plot of **a**, where $\Delta I/I_0$ is shown by the colour scale. In **b** we use interpolation to show a clearer picture on the frequency shifting at larger pump fluences.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Optical and structural changes in WTe_2 induced by 2.1-μm pump pulses. a, The transient reflectivity of the 800-nm probe gives a direct experimental probe to the electronic system. There is an abrupt change in $\Delta R/R$ right after the pump pulse arrival (within 5 ps). Afterwards, the $\Delta R/R$ signal remains finite and stable for longer than 50 ps. **b**, Bragg peak intensity changes probed by the electron beam. The intensity changes show oscillations that correspond to the interlayer shear-mode frequency of 0.24 THz, similar to the effect produced by the terahertz pump pulses discussed in the main text. **c**, Time-resolved SHG of WTe_2 at nanosecond time delay. Here, the pump pulse has a wavelength of 2.1 μm (polarized at 45° off the horizontal axis), the incident probe pulse has a wavelength of 800 nm, the crystal *a* axis is aligned horizontally and the SHG is detected at the 'S-in, P-out' configuration. This shows that the light-induced centrosymmetric phase lives for a few nanoseconds, or even tens of nanoseconds, which is consistent with the induced metastable phase discussed in the main text.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Lattice and electronic structure parameters for DFT calculations

(a)

	W(1)	W(2)	Te(1)	Te(2)	Te(3)	Te(4)
y	0.60062	0.03980	0.85761	0.64631	0.29845	0.20722
z	0.5	0.01522	0.65525	0.11112	0.85983	0.40387

(b)

	k_x	k_y	k_z	C
WP1	0.12195	0.03947	0	+1
WP2	0.12160	0.04510	0	-1

(c)

	a (Å)	b (Å)	c (Å)	d_1 (Å)	d_2 (Å)	d_3 (Å)	ν_{shear} (THz)
D3	3.4807	6.2806	13.9973	5.3104	3.8734	5.0137	0.15
D3-BJ	3.4841	6.2528	13.5910	5.2570	3.7169	4.7787	0.34
Experiment [5]	3.477	6.249	14.018	5.397	3.911	4.937	0.24

a, Wyckoff position parameters for the coordinates of two W and four Te atoms in the experimentally determined unit cell of WTe_2[16,42]. **b**, Coordinates of the two WPs that occur in the first quadrant of the $k_z = 0$ plane of the BZ of WTe_2. The last column indicates the Chern number (C) of the nodes. **c**, Calculated lattice and interlayer distance parameters of Td WTe_2 compared to experimental values. The last column shows the calculated phonon frequency (ν_{shear}) of the *b*-axis shear mode of interest. This table has been reproduced from ref. [16], Springer Nature.

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0808-5

Enzymatic assembly of carbon–carbon bonds via iron-catalysed sp^3 C–H functionalization

Ruijie K. Zhang[1], Kai Chen[1], Xiongyi Huang[1], Lena Wohlschlager[1,2], Hans Renata[1,3] & Frances H. Arnold[1]*

Although abundant in organic molecules, carbon–hydrogen (C–H) bonds are typically considered unreactive and unavailable for chemical manipulation. Recent advances in C–H functionalization technology have begun to transform this logic, while emphasizing the importance of and challenges associated with selective alkylation at a sp^3 carbon[1,2]. Here we describe iron-based catalysts for the enantio-, regio- and chemoselective intermolecular alkylation of sp^3 C–H bonds through carbene C–H insertion. The catalysts, derived from a cytochrome P450 enzyme in which the native cysteine axial ligand has been substituted for serine (cytochrome P411), are fully genetically encoded and produced in bacteria, where they can be tuned by directed evolution for activity and selectivity. That these proteins activate iron, the most abundant transition metal, to perform this chemistry provides a desirable alternative to noble-metal catalysts, which have dominated the field of C–H functionalization[1,2]. The laboratory-evolved enzymes functionalize diverse substrates containing benzylic, allylic or α-amino C–H bonds with high turnover and excellent selectivity. Furthermore, they have enabled the development of concise routes to several natural products. The use of the native iron-haem cofactor of these enzymes to mediate sp^3 C–H alkylation suggests that diverse haem proteins could serve as potential catalysts for this abiological transformation, and will facilitate the development of new enzymatic C–H functionalization reactions for applications in chemistry and synthetic biology.

Biological systems use a limited set of chemical strategies to form carbon–carbon (C–C) bonds during the construction of organic molecules[3]. Whereas many of these approaches rely on the manipulation of functional groups, certain enzymes—including members of the radical *S*-adenosylmethionine (SAM) family—can perform alkylation of sp^3 C–H bonds. This is a versatile strategy for structural diversification, as seen by its essential role in the biosynthesis of structurally varied natural products and cofactors[4–6]. However, known biological machineries for this transformation are limited to enzymes that transfer a methyl group[5,6] or conjugate a radical acceptor substrate[4,7] to specific molecules, with methylation as a common mode of sp^3 C–alkyl installation by radical SAM enzymes (Fig. 1a).

We sought to introduce a new enzymatic strategy for the alkylation of sp^3 C–H bonds. For our design, we drew inspiration from the most widely used biological C–H functionalization transformation: C–H oxygenation. Enzymes such as the cytochromes P450 accomplish C–H oxygenation using a haem cofactor; their activities rely on the activation of molecular oxygen for the controlled generation of a high-energy iron-oxo intermediate that is capable of selective insertion into a substrate C–H bond[8]. Analogously, we anticipated that the combination of a haem protein and a diazo compound would generate a protein-enclosed iron carbene species, and that this carbene could participate in a selective C–H insertion reaction with a second substrate (Fig. 1b). Although it has been shown that haem proteins are capable of performing carbene-transfer processes such as cyclopropanation and heteroatom–hydrogen bond insertions[9–11], their functionalization of sp^3 C–H bonds is yet to be achieved.

Metal carbene sp^3 C–H insertion in small-molecule catalysis, in particular intermolecular and stereoselective versions of this reaction, typically relies on transition-metal complexes based on rhodium[12], iridium[13] and other metals[14–16]. Artificial metalloproteins for carbene C–H insertion have been created by introducing an iridium-porphyrin into variants of apo haem proteins[17]. Although rare, there are a few examples of iron carbene sp^3 C–H insertion. The iron-catalysed examples use high temperatures (for example 80 °C)[18], are stoichiometric[19] or are restricted to intramolecular reactions[20], which suggests there is a high activation energy barrier to the insertion of an iron carbene into a C–H bond. However, because the protein framework of an enzyme can impart substantial rate enhancements to reactions[21] and even confer activity to an otherwise unreactive cofactor[22], we surmised that directed evolution could reconfigure a haem protein to overcome the barrier for

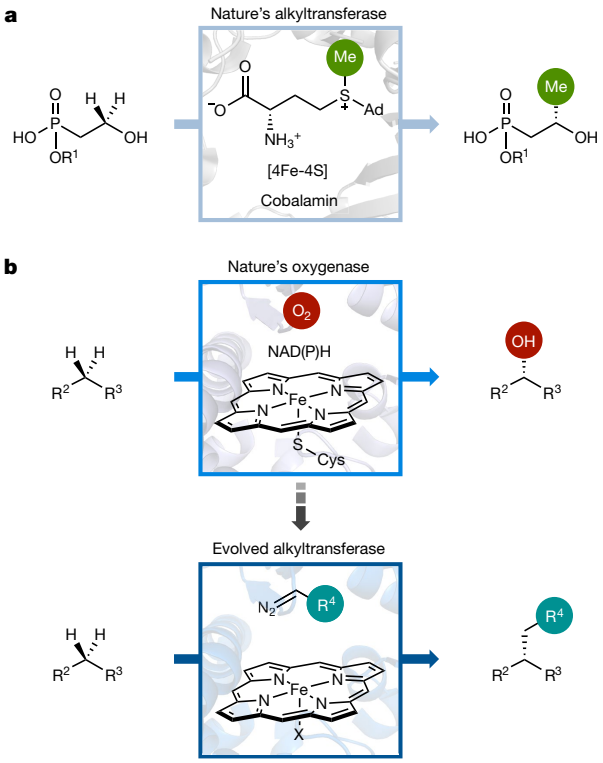


Fig. 1 | Enzymatic C–H functionalization systems. a, Methylation catalysed by cobalamin-dependent radical SAM enzymes, as illustrated by Fom3 in fosfomycin biosynthesis[6]. **b**, Oxygenation catalysed by cytochrome P450 monooxygenase (top) and the proposed alkylation reaction achieved under haem protein catalysis (bottom). Structural illustrations are adapted from Protein Data Bank (PDB) ID 5UL4 (radical SAM enzyme) and PDB 2IJ2 (cytochrome $P450_{BM3}$). Ad, adenosyl; Cys, cysteine; R, organic group; X, amino acid.

[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, USA. [2]Present address: Institute of Food Technology, University of Natural Resources and Life Sciences, Vienna, Austria. [3]Present address: Department of Chemistry, The Scripps Research Institute, Jupiter, FL, USA. *e-mail: frances@cheme.caltech.edu

© 2019 Springer Nature Limited. All rights reserved.

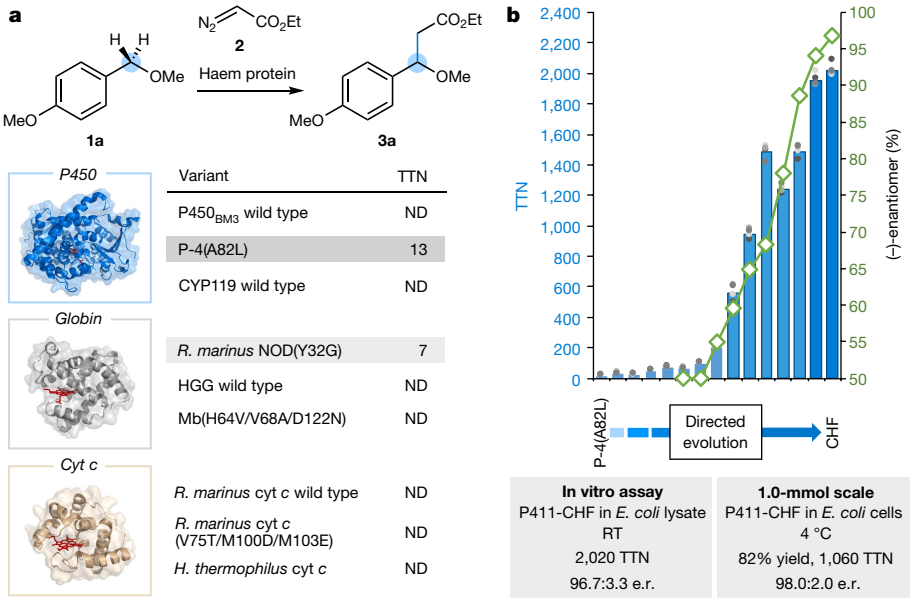


Fig. 2 | Haem-protein-catalysed sp^3 C–H alkylation. a, A selected subset of haem proteins that were tested for promiscuous C–H alkylation activity. Structural illustrations are of the following superfamily members with the haem cofactor shown as red sticks: cytochrome $P450_{BM3}$ (top, PDB 2IJ2), sperm whale myoglobin (middle, PDB 1A6K) and *R. marinus* cytochrome *c* (bottom, PDB 3CP5). cyt *c*, cytochrome *c*; HGG, Hell's Gate globin; *H. thermophilus*, *Hydrogenobacter thermophilus*; Mb, sperm whale myoglobin; ND, not detected. **b**, Directed evolution of a cytochrome P411 for enantioselective C–H alkylation (reaction shown in **a**). Bars represent mean TTN values averaged over four reactions (performed from two independent cell cultures, each used for duplicate reactions); each TTN data point is shown as a grey dot. Enantioselectivity data are represented by green diamonds. Unless otherwise indicated, reaction conditions were as follows: haem protein in *E. coli* whole cells (optical density at 600 nm, OD_{600}, of 30) (**a**) or in clarified *E. coli* lysate (**b**), 10 mM substrate **1a**, 10 mM ethyl diazoacetate, 5 vol% EtOH in M9-N buffer at room temperature (RT) under anaerobic conditions for 18 h; see Supplementary Information for conditions of the 1.0-mmol reaction in **b**. Reactions performed with lysate contain 1 mM $Na_2S_2O_4$. TTN is defined as the amount of indicated product divided by haem protein as measured by the haemochrome assay. See Supplementary Information for the complete list of haem proteins tested and detailed experimental procedures.

the iron carbene C–H insertion reaction and acquire this new function (Fig. 1b).

In initial studies, we tested a panel of 78 haem proteins that included variants of cytochromes P450, cytochromes *c* and globin homologues. The haem proteins in whole *Escherichia coli* cells were combined with *p*-methoxybenzyl methyl ether (**1a**) and ethyl diazoacetate (**2**) at room temperature under anaerobic conditions; the resulting reactions were analysed for the formation of C–H alkylation product **3a** (Fig. 2a; see Supplementary Information for the complete list of haem proteins tested). Haem proteins from two superfamilies were found to show low levels of this promiscuous activity, establishing the possibility of creating C–H alkylation enzymes with very different protein architectures. Among the proteins tested were variants of cytochrome $P450_{BM3}$ from *Bacillus megaterium* in which the axial cysteine ligand is substituted for serine, known as cytochrome P411s[31]. We found that one of these variants, P-4(A82L)[22], which differs from the wild type by 18 mutations, provided **3a** with a total turnover number (TTN) of 13. In addition, nitric oxide dioxygenase from *Rhodothermus marinus* with a tyrosine-to-glycine mutation at position 32 (*R. marinus* NOD(Y32G)) catalysed the reaction with 7 TTN. A second alkane substrate, 4-ethylanisole (**1i**), was also accepted by the nascent C–H alkylation enzymes, albeit with lower turnover numbers (Supplementary Table 2). The haem cofactor alone (iron protoporphyrin IX) or in the presence of bovine serum albumin was inactive (Supplementary Tables 1 and 2).

With P411 P-4(A82L) as the starting template, sequential rounds of site-saturation mutagenesis and screening in whole *E. coli* cells were performed to identify increasingly active and enantioselective biocatalysts for C–H alkylation. Amino acid residues chosen for mutagenesis included those that line the active site pocket, reside on loops and other flexible regions of the protein, or possess a nucleophilic side chain[23]. Improved variants were subsequently evaluated in reactions using clarified *E. coli* lysate with *p*-methoxybenzyl methyl ether (**1a**) and 4-ethylanisole (**1i**) (Fig. 2b and Supplementary Fig. 1). Five rounds of mutagenesis and screening yielded variant P411-gen6, which furnished product **3a** with 60 TTN. Unlike the native monooxygenase activity, the C–H alkylation process does not require reducing equivalents from the FAD and FMN domains of these enzymes. Surmising that these domains may not be needed for the C–H alkylation reaction, we performed systematic truncations of P411-gen6 to determine the minimally sufficient domain(s) for retaining catalytic activity. Notably, removal of the FAD domain, which contains 37% of the amino acids in the full-length protein, created an enzyme with higher C–H alkylation activity: P411ΔFAD-gen6 delivers **3a** with 100 TTN, a 1.7-fold increase in TTN compared with P411-gen6 (Supplementary Fig. 2). This indicates that the FAD domain may have (negative) allosteric effects on the C–H alkylation activity. Further studies with these truncated enzymes revealed that they could be used in whole *E. coli* cells, in clarified *E. coli* cell lysate and as purified proteins (Supplementary Table 3). Eight additional rounds of mutagenesis and screening yielded P411-CHF (P411ΔFAD C–H functionalization enzyme; for the full list of changes, see Supplementary Information).

P411-CHF displays a 140-fold improvement in activity over P-4(A82L) and delivers **3a** with excellent stereoselectivity (2,020 TTN, 96.7: 3.3 enantiomeric ratio (e.r.) using clarified *E. coli* lysate). Subsequent studies showed that the stereoselectivity could be improved by conducting the reaction at lower temperature (for example, 4 °C) without substantial change to TTN (Supplementary Table 4). Enzymatic C–H alkylation can be performed on a millimole scale: using 1.0 mmol substrate **1a**, *E. coli* harbouring P411-CHF at 4 °C furnished **3a** in 82% isolated yield, 1,060 TTN, and 98.0: 2.0 e.r. (Fig. 2b). Preliminary mechanistic investigations were pursued to investigate the nature of the C–H insertion step. Independent initial rates measured for reactions with substrate **1a** or deuterated substrate **1a-d_2** revealed a normal kinetic isotope effect of 5.1 for C–H alkylation catalysed by P411-CHF, suggesting that C–H insertion is rate-determining (Supplementary Fig. 5).

Using *E. coli* harbouring P411-CHF, we assayed a range of benzylic substrates for coupling with ethyl diazoacetate (Fig. 3). Both electron-rich and electron-deficient functionalities on the aromatic

© 2019 Springer Nature Limited. All rights reserved.

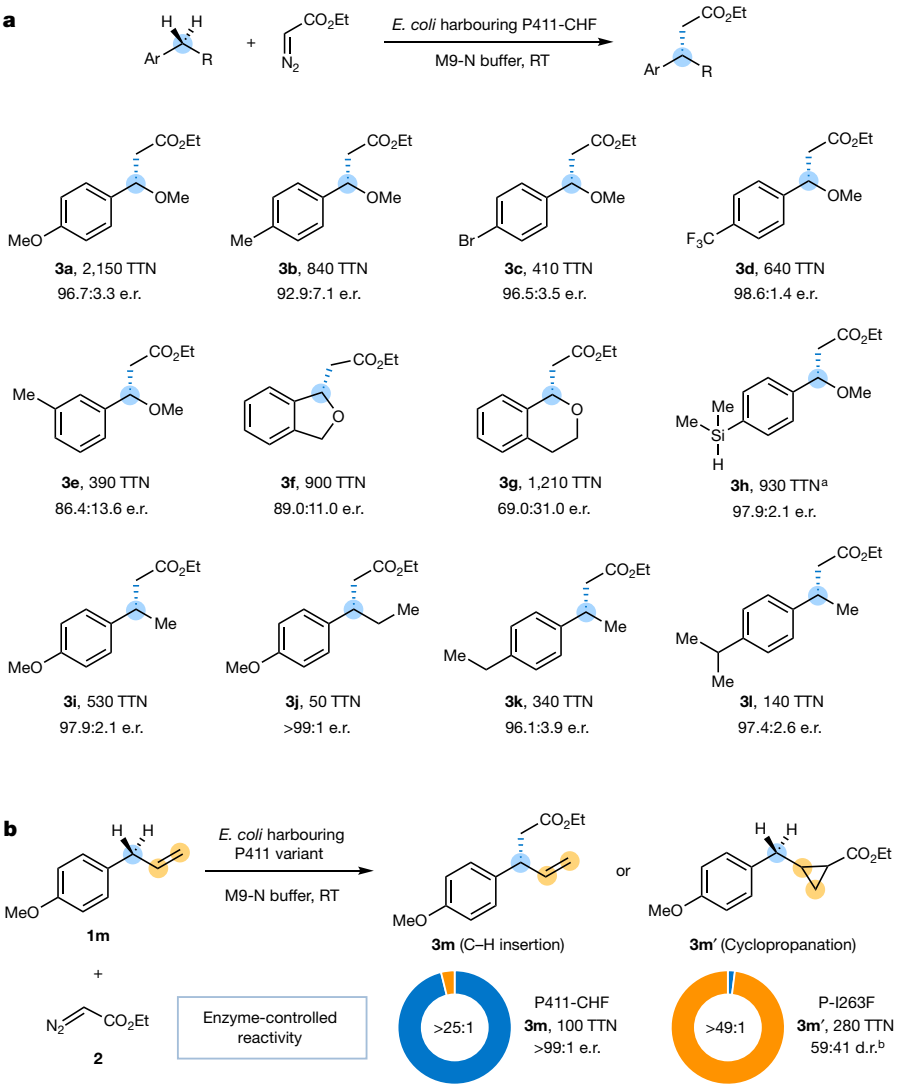


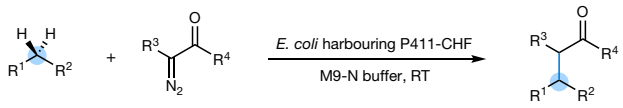
Fig. 3 | Substrate scope for benzylic C–H alkylation with P411-CHF. **a**, Experiments were performed using *E. coli* expressing cytochrome P411-CHF ($OD_{600} = 30$) with 10 mM substrate **1a**–**1l** and 10 mM ethyl diazoacetate at room temperature under anaerobic conditions for 18 h; reported TTNs are the average of four reactions (performed from two independent cell cultures, each used for duplicate reactions). See Supplementary Fig. 12 for the full list of alkane substrates. [a]Si–H insertion product **3h′** is also observed (Supplementary Fig. 7). **b**, Reaction selectivity for carbene C–H insertion or cyclopropanation can be controlled by the protein scaffold. Experiments were performed as in **a** using the indicated P411 variant. [b]Diastereomeric ratio (d.r.) is given as *cis*:*trans*; e.r. was not determined.

ring are well-tolerated (**3a**–**3e**, **3h**); cyclic substrates are also suitable coupling partners (**3f**, **3g**). The functionalization of alkyl benzenes at secondary benzylic sp^3 C–H bonds was found to be successful (**3i**–**3l**). Notably, in the biotransformation of substrate **1l**, which contains both tertiary and secondary benzylic C–H bonds, P411-CHF preferentially functionalizes the secondary position despite its higher C–H bond dissociation energy. The carbene intermediate derived from ethyl diazoacetate belongs to the acceptor-only class. Compared to the more widely used donor/acceptor carbenes, acceptor-only intermediates are more electrophilic, and as a result selective reactions with this carbene class are still a major challenge for small-molecule catalysts[13,16]. Our results show that P411-CHF can control this highly reactive intermediate to furnish the desired sp^3 C–H alkylation products, and does so with high enantioselectivity.

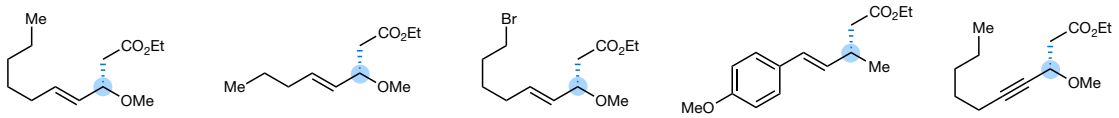
Enzymes can exhibit excellent reaction selectivity arising from their ability to form multiple interactions with substrates and intermediates throughout a reaction cycle. We proposed that the protein scaffold could be tuned to create complementary enzymes that can access different reaction outcomes available to a substrate. When P411-CHF was challenged with 4-allylanisole (**1m**)—a substrate that can undergo both C–H alkylation and cyclopropanation—we observed that C–H alkylation product **3m** dominated, with selectivity > 25:1 (Fig. 3b, Supplementary Fig. 6). By contrast, a related full-length P411 variant P-I263F, containing 13 mutations in the haem domain relative to P411-CHF, catalysed only the formation of cyclopropane product **3m′**. Additionally, despite the established reactivity of silanes with iron carbenes[10], P411-CHF delivered C–H alkylation product **3h** when substrate **1h** was used in the reaction (Si–H insertion product **3h′** was also observed, but its formation may not be catalysed by P411-CHF, Supplementary Fig. 7). Reaction with P-I263F, by contrast, provided only the Si–H insertion product. These examples demonstrate an exceptional feature of macromolecular enzymes: different products can be obtained simply by changing the amino acid sequence of the protein catalyst.

Enzymatic C–H alkylation is not limited to the functionalization of benzylic C–H bonds. Structurally dissimilar molecules containing allylic or propargylic C–H bonds are excellent substrates for this chemistry (Fig. 4a). In contrast to **1a**–**1m**, which contain a rigid benzene ring, compounds **4a**–**4c** and **4e** feature flexible linear alkyl chains. Their successful enantioselective alkylation suggests that the enzyme active site can accommodate substrate conformational flexibility while enforcing a favoured substrate orientation relative to the carbene intermediate.

© 2019 Springer Nature Limited. All rights reserved.



a Allylic and propargylic substrates



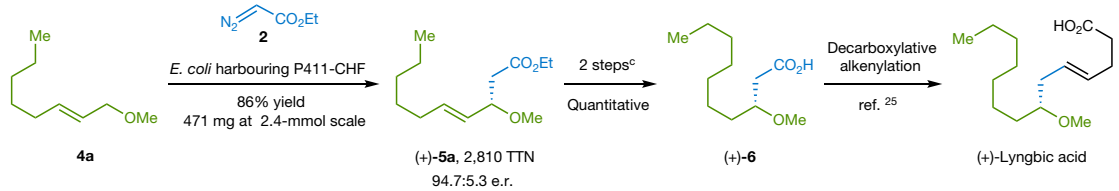
5a, 3,750 TTN
93.6:6.4 e.r.

5b, 950 TTN
94.7:5.3 e.r.

5c, 2,760 TTN
96.3:3.7 e.r.

5d, 70 TTN[a]
97.0:3.0 e.r.

5e, 190 TTN[b]
99.0:1.0 e.r.



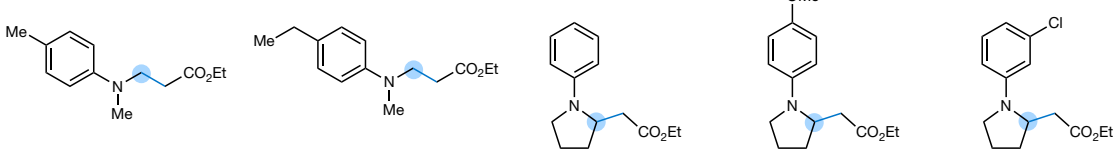
4a

(+)-**5a**, 2,810 TTN
94.7:5.3 e.r.

(+)-**6**

(+)-Lyngbic acid

b Alkyl amines



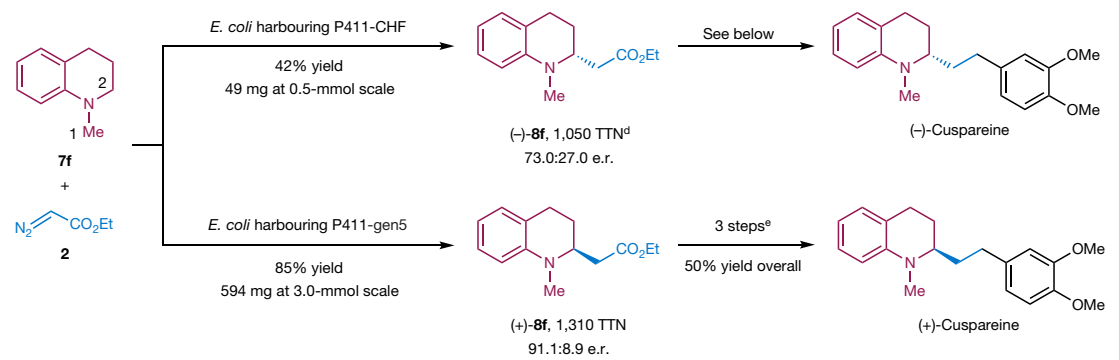
8a, 2,330 TTN

8b, 2,230 TTN

8c, 2,030 TTN
82.8:17.2 e.r.

8d, 2,150 TTN
83.7:16.3 e.r.

8e, 500 TTN
90.3:9.7 e.r.



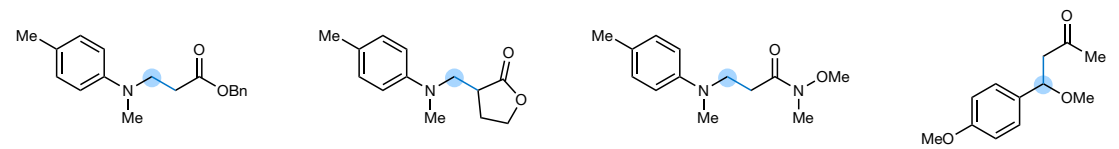
(–)-**8f**, 1,050 TTN[d]
73.0:27.0 e.r.

(–)-Cuspareine

(+)-**8f**, 1,310 TTN
91.1:8.9 e.r.

(+)-Cuspareine

c Alternative diazos



10a, 360 TTN

10b, 280 TTN[f]
78.0:22.0 e.r.

10c, 500 TTN[f]

10d, 150 TTN
71.0:29.0 e.r.

Fig. 4 | Application of P411 enzymes for sp^3 C–H alkylation. a, Allylic and propargylic C–H alkylation. Unless otherwise indicated, experiments were performed using *E. coli* expressing cytochrome P411-CHF with 10 mM substrate **4a**–**4e** and 10 mM ethyl diazoacetate; reported TTNs are the average of four reactions (performed from two independent cell cultures, each used for duplicate reactions). [a]TTN was calculated on the basis of isolated yield from a reaction performed on a 0.25-mmol scale. [b]The cyclopropene product was also observed (Supplementary Fig. 8). [c]Hydrogenation, followed by hydrolysis. **b**, Enzymatic alkylation of substrates containing α-amino C–H bonds. Unless otherwise indicated, reactions were performed on a 0.5-mmol scale using *E. coli* expressing cytochrome P411-CHF with substrates **7a**–**7f** and ethyl diazoacetate; TTNs were calculated on the basis of isolated yields of products shown. [d]Isolated in 9:1 r.r. for **8f**:**8f′** determined by ^{1}H nuclear magnetic resonance analysis; TTN is reported for the sum of the regioisomers. [e]Reduction, halogen exchange and Suzuki–Miyaura cross-coupling. **c**, Enzymatic C–H alkylation with alternative diazo reagents. Unless otherwise indicated, reactions were performed on a 0.5-mmol scale using *E. coli* expressing cytochrome P411-CHF with coupling partner **1a** or **7a** and diazo compounds **9a**–**9d**; TTNs were calculated on the basis of isolated yields of products shown. [f]Variant P411-IY(T327I) was used. See Supplementary Information for the complete list of substrates (Supplementary Figs. 12, 13), information about enzyme variants, and full experimental details.

© 2019 Springer Nature Limited. All rights reserved.

To demonstrate the utility of this biotransformation, we applied the methodology to the formal synthesis of lyngbic acid (Fig. 4a). Marine cyanobacteria incorporate this versatile biomolecule into members of the malyngamide family of natural products; likewise, total-synthesis approaches to malyngamides typically access lyngbic acid as a strategic intermediate en route to the target molecules[24]. Using *E. coli* harbouring P411-CHF, intermediate **5a** was produced on a 2.4-mmol scale in 86% isolated yield, 2,810 TTN, and 94.7:5.3 e.r. Subsequent hydrogenation and hydrolysis provided (*R*)-(+)-**6** in quantitative yield, which can be transformed to (*R*)-(+)-lyngbic acid by decarboxylative alkenylation[25].

As part of our investigation into the substrate scope of the reaction, we challenged P411-CHF with alkyl amine compounds. Compounds of this type are typically challenging substrates for C–H functionalization methods because the amine functionality may coordinate to and inhibit the catalyst or undergo undesirable side reactions (for example, ylide formation and its associated rearrangements)[26]. Using **7a** or **7b**, substrates that have both benzylic C–H bonds and α-amino C–H bonds, P411-CHF delivered the corresponding β-amino ester product with high efficiency (**8a** and **8b**, Fig. 4b). Notably, benzylic C–H insertion was either not observed (with **7a**, Supplementary Fig. 9) or was considerably suppressed (with **7b**, Supplementary Fig. 10), despite the typically lower bond dissociation energies of benzylic C–H bonds compared to α-amino C–H bonds. Additionally, *N*-aryl pyrrolidines (**7c**–**7e**) were found to be excellent substrates and were selectively alkylated at the α-amino sp^3 position. Using P411-CHF, the sp^3 C–H alkylation of **7c** outcompetes a Friedel–Crafts type reaction on the aryl ring, which is a favourable process with other carbene-transfer systems[27,28]. Furthermore, alkylation product **8d** offers a conceivable strategy for the synthesis of β-homoproline, a motif that has been investigated for medicinal chemistry applications[29].

Given that P411-CHF alkylates both primary and secondary α-amino C–H bonds, we investigated whether the enzyme could be selective for one of these positions. Using *N*-methyl tetrahydroquinoline **7f** as the alkane substrate, P411-CHF afforded β-amino ester products with 1,050 TTN and a 9:1 ratio of regioisomers (C2:C1, and 73.0:27.0 e.r. for (−)-**8f**) (Fig. 4b). The tetrahydroquinoline ring is a prevalent structural motif in natural products and bioactive molecules[30], and its selective functionalization could provide a concise strategy for the synthesis of alkaloids. To improve the selectivity for the alkylation of **7f**, we tested variants along the evolutionary lineage from P-4(A82L) to P411-CHF. We found that, compared with P411-CHF, P411-gen5 showed even better regioselectivity and the opposite stereochemical preference for C–C bond formation. In a reaction on 3.0-mmol scale, *E. coli* harbouring P411-gen5 delivered (+)-**8f** in 85% yield with excellent selectivity (1,310 TTN, >50:1 regiomeric ratio (r.r.), 91.1:8.9 e.r.). In only a few steps, the enzymatic product was successfully transformed to the alkaloid (*R*)-(+)-cuspareine[30] (Fig. 4b).

Finally, we explored the introduction of different alkyl groups. Using different diazo reagents, enzymatic C–H alkylation can diversify one alkane substrate, such as **7a**, to several products (**10a**–**10c** in Fig. 4c and Supplementary Fig. 11). The diazo substrate scope extends beyond ester-based reagents: Weinreb amide diazo compound **9c** and diazoketone **9d** were found to participate in enzymatic C–H alkylation to furnish products **10c** and **10d**, respectively. Additional substitution at the α-position of the carbene, however, is generally not well-tolerated by P411-CHF and the current related enzymes. With the exception of **10b**, reactions using disubstituted carbene reagents did not yield appreciable amounts of desired products (Supplementary Fig. 11).

This study demonstrates that a cytochrome P450 can acquire the ability to construct C–C bonds from sp^3 C–H bonds, and that the activity and selectivity of the reaction can be greatly enhanced using directed evolution. Nature provides a huge collection of possible alternative starting points for expanding the scope of this reaction even further and for achieving other selectivities. The cytochrome P450 superfamily can access an immense set of organic molecules for its native oxygenation chemistry; we foresee that P411-derived enzymes and other natural haem protein diversity can be leveraged to generate families of C–H alkylation enzymes that emulate the scope and selectivity of nature's C–H oxygenation catalysts.

Data availability

All relevant data are provided in Supplementary Information. Any additional information is available from the corresponding author on request.

Received: 18 July 2018; Accepted: 1 November 2018;
Published online 19 December 2018.

1. Hartwig, J. F. & Larsen, M. A. Undirected, homogeneous C–H bond functionalization: challenges and opportunities. *ACS Cent. Sci.* **2**, 281–292 (2016).
2. Saint-Denis, T. G., Zhu, R.-Y., Chen, G., Wu, Q.-F. & Yu, J.-Q. Enantioselective C(sp^3)–H bond activation by chiral transition metal catalysts. *Science* **359**, eaao4798 (2018).
3. Frey, P. A. & Hegeman, A. D. *Enzymatic Reaction Mechanisms* Ch. 14 (Oxford Univ. Press, New York, 2007).
4. Yokoyama, K. & Lilla, E. A. C–C bond forming radical SAM enzymes involved in the construction of carbon skeletons of cofactors and natural products. *Nat. Prod. Rep.* **35**, 660–694 (2018).
5. Bauerle, M. R., Schwalm, E. L. & Booker, S. J. Mechanistic diversity of radical *S*-adenosylmethionine (SAM)-dependent methylation. *J. Biol. Chem.* **290**, 3995–4002 (2015).
6. McLaughlin, M. I. & van der Donk, W. A. Stereospecific radical-mediated B_{12}-dependent methyl transfer by the fosfomycin biosynthesis enzyme Fom3. *Biochemistry* **57**, 4967–4971 (2018).
7. Shisler, K. A. & Broderick, J. B. Glycyl radical activating enzymes: structure, mechanism, and substrate interactions. *Arch. Biochem. Biophys.* **546**, 64–71 (2014).
8. Poulos, T. L. Heme enzyme structure and function. *Chem. Rev.* **114**, 3919–3962 (2014).
9. Coelho, P. S., Brustad, E. M., Kannan, A. & Arnold, F. H. Olefin cyclopropanation via carbene transfer catalyzed by engineered cytochrome P450 enzymes. *Science* **339**, 307–310 (2013).
10. Kan, S. B. J., Lewis, R. D., Chen, K. & Arnold, F. H. Directed evolution of cytochrome c for carbon–silicon bond formation: bringing silicon to life. *Science* **354**, 1048–1051 (2016).
11. Brandenberg, O. F., Fasan, R. & Arnold, F. H. Exploiting and engineering hemoproteins for abiological carbene and nitrene transfer reactions. *Curr. Opin. Biotechnol.* **47**, 102–111 (2017).
12. Liao, K. et al. Site-selective and stereoselective functionalization of non-activated tertiary C–H bonds. *Nature* **551**, 609–613 (2017).
13. Weldy, N. M. et al. Iridium(III)-bis(imidazolinyl)phenyl catalysts for enantioselective C–H functionalization with ethyl diazoacetate. *Chem. Sci.* **7**, 3142–3146 (2016).
14. Wang, Y., Wen, X., Cui, X. & Zhang, X. P. Enantioselective radical cyclization for construction of 5-membered ring structures by metalloradical C–H alkylation. *J. Am. Chem. Soc.* **140**, 4792–4796 (2018).
15. Caballero, A. et al. Silver-catalyzed C–C bond formation between methane and ethyl diazoacetate in supercritical CO_2. *Science* **332**, 835–838 (2011).
16. Wu, W.-T., Yang, Z.-P. & You, S.-L. in *Asymmetric Functionalization of C–H Bonds* (ed. You, S.-L.) Ch. 1 (Royal Society of Chemistry, Cambridge, 2015).
17. Key, H. M., Dydio, P., Clark, D. S. & Hartwig, J. F. Abiological catalysis by artificial haem proteins containing noble metals in place of iron. *Nature* **534**, 534–537 (2016).
18. Mbuvi, H. M. & Woo, L. K. Catalytic C–H insertions using iron(III) porphyrin complexes. *Organometallics* **27**, 637–645 (2008).
19. Li, Y., Huang, J.-S., Zhou, Z.-Y., Che, C.-M. & You, X.-Z. Remarkably stable iron porphyrins bearing nonheteroatom-stabilized carbene or (alkoxycarbonyl) carbenes: isolation, X-ray crystal structures, and carbon atom transfer reactions with hydrocarbons. *J. Am. Chem. Soc.* **124**, 13185–13193 (2002).
20. Griffin, J. R., Wendell, C. I., Garwin, J. A. & White, M. C. Catalytic C(sp^3)–H alkylation via an iron carbene intermediate. *J. Am. Chem. Soc.* **139**, 13624–13627 (2017).
21. Herschlag, D. & Natarajan, A. Fundamental challenges in mechanistic enzymology: progress toward understanding the rate enhancements of enzymes. *Biochemistry* **52**, 2050–2067 (2013).
22. Prier, C. K., Zhang, R. K., Buller, A. R., Brinkmann-Chen, S. & Arnold, F. H. Enantioselective, intermolecular benzylic C–H amination catalysed by an engineered iron-haem enzyme. *Nat. Chem.* **9**, 629–634 (2017).
23. Renata, H. et al. Identification of mechanism-based inactivation in P450-catalyzed cyclopropanation facilitates engineering of improved enzymes. *J. Am. Chem. Soc.* **138**, 12527–12533 (2016).
24. Zhang, J.-T. et al. Total synthesis of malyngamides K, L, and 5″-*epi*-C and absolute configuration of malyngamide L. *J. Org. Chem.* **76**, 3946–3959 (2011).
25. Edwards, J. T. et al. Decarboxylative alkenylation. *Nature* **545**, 213–218 (2017).
26. He, J., Hamann, L. G., Davies, H. M. L. & Beckwith, R. E. J. Late-stage C–H functionalization of complex alkaloids and drug molecules via intermolecular rhodium-carbenoid insertion. *Nat. Commun.* **6**, 5943 (2015).
27. Yang, J.-M., Cai, Y., Zhu, S.-F. & Zhou, Q.-L. Iron-catalyzed arylation of α-aryl-α-diazoesters. *Org. Biomol. Chem.* **14**, 5516–5519 (2016).

© 2019 Springer Nature Limited. All rights reserved.

28. Xu, B., Li, M.-L., Zuo, X.-D., Zhu, S.-F. & Zhou, Q.-L. Catalytic asymmetric arylation of α-aryl-α-diazoacetates with aniline derivatives. *J. Am. Chem. Soc.* **137**, 8700–8703 (2015).
29. Cardillo, G., Gentilucci, L., Qasem, A. R., Sgarzi, F. & Spampinato, S. Endomorphin-1 analogues containing β-proline are μ-opioid receptor agonists and display enhanced enzymatic hydrolysis resistance. *J. Med. Chem.* **45**, 2571–2578 (2002).
30. Sridharan, V., Suryavanshi, P. A. & Menéndez, J. C. Advances in the chemistry of tetrahydroquinolines. *Chem. Rev.* **111**, 7157–7259 (2011).
31. Coelho, P. S. et al. A serine-substituted P450 catalyzes highly efficient carbene transfer to olefins *in vivo*. *Nat. Chem. Biol.* **9**, 485–487 (2013).

Acknowledgements This work was supported by the National Science Foundation (NSF), Division of Molecular and Cellular Biosciences (grant MCB-1513007). R.K.Z. acknowledges support from the NSF Graduate Research Fellowship (grant DGE-1144469) and the Donna and Benjamin M. Rosen Bioengineering Center. X.H. is supported by a Ruth L. Kirschstein National Institutes of Health Postdoctoral Fellowship (grant F32GM125231). L.W. received support from the Austrian Marshall Plan Foundation. We thank A. Z. Zhou for experimental assistance; N. W. Goldberg, S. C. Hammer, K. E. Hernandez, Z. Jia, A. M. Knight, G. Kubik, R. D. Lewis, C. K. Prier, D. K. Romney and J. Zhang for discussions; S. Virgil, M. Shahgholi and D. VanderVelde for analytical support; and B. Stoltz for use of polarimeter and gas chromatography equipment.

Reviewer information *Nature* thanks B. de Bruin, N. Turner and T. Ward for their contribution to the peer review of this work.

Author contributions R.K.Z. designed the overall research with F.H.A. providing guidance. R.K.Z. and H.R. designed and conducted the initial screening of haem proteins; R.K.Z. and L.W. performed the directed evolution experiments. R.K.Z., K.C. and X.H. designed and performed the substrate scope studies. R.K.Z. and F.H.A. wrote the manuscript with input from all authors.

Competing interests A provisional patent application has been filed through the California Institute of Technology based on the results presented here.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0808-5.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to F.H.A.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0800-0

Greenland melt drives continuous export of methane from the ice–sheet bed

Guillaume Lamarche-Gagnon[1]*, Jemma L. Wadham[1], Barbara Sherwood Lollar[2], Sandra Arndt[3], Peer Fietzek[4], Alexander D. Beaton[5], Andrew J. Tedstone[1], Jon Telling[6], Elizabeth A. Bagshaw[7], Jon R. Hawkings[1,8,9], Tyler J. Kohler[10], Jakub D. Zarsky[10], Matthew C. Mowlem[5], Alexandre M. Anesio[11] & Marek Stibal[10]

Ice sheets are currently ignored in global methane budgets[1,2]. Although ice sheets have been proposed to contain large reserves of methane that may contribute to a rise in atmospheric methane concentration if released during periods of rapid ice retreat[3,4], no data exist on the current methane footprint of ice sheets. Here we find that subglacially produced methane is rapidly driven to the ice margin by the efficient drainage system of a subglacial catchment of the Greenland ice sheet. We report the continuous export of methane-supersaturated waters ($CH_{4(aq)}$) from the ice-sheet bed during the melt season. Pulses of high $CH_{4(aq)}$ concentration coincide with supraglacially forced subglacial flushing events, confirming a subglacial source and highlighting the influence of melt on methane export. Sustained methane fluxes over the melt season are indicative of subglacial methane reserves that exceed methane export, with an estimated 6.3 tonnes (discharge-weighted mean; range from 2.4 to 11 tonnes) of $CH_{4(aq)}$ transported laterally from the ice-sheet bed. Stable-isotope analyses reveal a microbial origin for methane, probably from a mixture of inorganic and ancient organic carbon buried beneath the ice. We show that subglacial hydrology is crucial for controlling methane fluxes from the ice sheet, with efficient drainage limiting the extent of methane oxidation[5] to about 17 per cent of methane exported. Atmospheric evasion is the main methane sink once runoff reaches the ice margin, with estimated diffusive fluxes (4.4 to 28 millimoles of CH_4 per square metre per day) rivalling that of major world rivers[6]. Overall, our results indicate that ice sheets overlie extensive, biologically active methanogenic wetlands and that high rates of methane export to the atmosphere can occur via efficient subglacial drainage pathways. Our findings suggest that such environments have been previously underappreciated and should be considered in Earth's methane budget.

The role of ice sheets in the global methane cycle depends on the ability (thermogenic or microbial) of subglacial environments to produce large quantities of methane (for example, as hydrates)[3,4,7], as well as the mechanisms responsible for methane export to the ice margin and subsequent release to the atmosphere. Subglacial methane hydrates have been suggested to currently exist beneath the Antarctic Ice Sheet in large enough quantities to raise atmospheric methane concentrations if released rapidly during deglaciation[4]. However, recent research has revealed the presence of active methane-oxidizing communities in subglacial ecosystems, suggesting the possibility of an efficient methane buffer by an active biological sink[5,8]. There is also ambiguity in the palaeo-record. New ice-core data suggest that geological methane (for example, from permafrost but also potentially of ice-sheet origin) had little effect on atmospheric methane concentrations over the Younger Dryas–Preboreal transition[9]; but previous estimates do suggest large subglacial methane releases from retreating palaeo-ice sheets of the Northern Hemisphere following the onset of the last deglaciation[10]. Confounding scenarios on the potency of sub-ice-sheet methane mostly result from the scarcity of empirical data, which are limited to point measurements in ice cores[11–13], Greenland marginal streams[5] and an Antarctic subglacial lake[8].

Here we provide direct evidence from the Greenland ice sheet (GrIS) for the existence of large subglacial methane reserves, where production is not offset by local sinks and there is net export of methane to the atmosphere during the summer melt season. We focus on a 600-km^2 catchment of the GrIS that has been extensively studied over the last decade, both in terms of ice dynamics and subglacial geochemistry (Supplementary Information 1a). Between 19 May and 13 July 2015, we deployed a CONTROS HydroC CH_4 sensor[14] (Kongsberg Maritime Contros, Germany) at a distance <2 km from the ice margin in the proglacial river of the Leverett Glacier (LG) (Supplementary Information 1a; Extended Data Fig. 1)[15,16]. Manual measurements supported sensor readings and CH_4 stable-isotope analyses ($\delta^{13}C$ and $\delta^{2}H$), and 16S rRNA gene-sequence data from LG runoff were used to infer the methane origin. A one-dimensional reaction-transport model was further applied to test for the possibility of hydrate formation beneath the ice in the catchment. Features of the study area suggest that results obtained are probably applicable to other ice-sheet catchments (Supplementary Information 1a) and are informative on a global scale, serving as a first-step assessment of subglacial methane contribution to present-day methane budgets.

Sensor measurements revealed that LG runoff was supersaturated in methane with respect to the atmosphere over the entire monitoring period (mean concentration of about 271 nM, compared with an atmospheric equilibrium concentration of about 4.5 nM) (Fig. 1). This is consistent with the high concentrations (up to about 24 μM) of methane detected in the basal regions of the GRIP, GISP2 and NGRIP ice cores[11–13], in marginal runoff from a small neighbouring Greenland glacier (about 3–83 μM)[5], and during experimental incubations of Greenland subglacial sediment[17]. Stepwise increases in methane concentrations closely followed the seasonal evolution of the subglacial drainage system, indicating the crucial role of hydrology in controlling methane export from the ice sheet. Clear differences in $CH_{4(aq)}$ concentrations were observed between (1) the early part of the season, during times of very low discharge, when the subglacial portal was completely ice-sealed and methane concentrations were low (mean concentration of about 64 nM) (Fig. 1; Supplementary Information 2b), (2) the emergence of a subglacial upwelling through the river ice in front of the LG on 1 June, which released methane-enriched waters stored over winter from the ice margin (mean concentration of about 4 μM before the melt season; see Supplementary Information 1b, Extended Data Fig. 1) and (3) the later season (from 19 June onwards), with elevated $CH_{4(aq)}$ concentrations (pulses) coincident with a series of four subglacial

[1]School of Geographical Sciences, University of Bristol, Bristol, UK. [2]Department of Earth Sciences, University of Toronto, Toronto, Ontario, Canada. [3]Department of Geoscience, Environment and Society, Université Libre de Bruxelles, Brussels, Belgium. [4]Kongsberg Maritime Contros GmbH, Kiel, Germany. [5]National Oceanography Centre, Southampton, UK. [6]School of Natural and Environmental Sciences, Newcastle University, Newcastle, UK. [7]School of Earth and Ocean Sciences, Cardiff University, Cardiff, UK. [8]National High Magnetic Field Lab and Earth, Ocean and Atmospheric Sciences, Florida State University, Tallahassee, FL, USA. [9]German Research Centre for Geosciences GFZ, Potsdam, Germany. [10]Department of Ecology, Faculty of Science, Charles University, Prague, Czechia. [11]Department of Environmental Sciences, Aarhus University, Roskilde, Denmark. *e-mail: guillaume.lg@bristol.ac.uk

© 2019 Springer Nature Limited. All rights reserved.

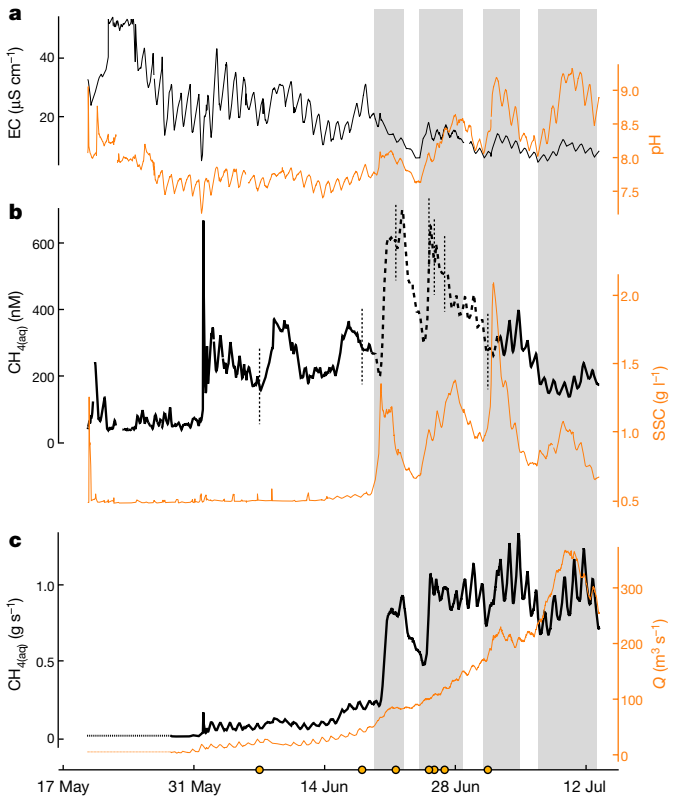


Fig. 1 | Geochemical time series of the LG proglacial river. a, Electrical conductivity (EC) and pH. **b**, $CH_{4(aq)}$ (data collected with the HydroC sensor) and suspended sediment concentrations (SSC). The dashed section corresponds to times when the HydroC sensor exhibited longer response times (see Supplementary Information 2a, Extended Data Fig. 2). Orange dots and vertical dashed lines indicate the sampling time of waters used for stable-isotope analysis (see Extended Data Table 2). **c**, $CH_{4(aq)}$ lateral flux and discharge (Q). The first data points, measured on 28 May, are extended to the first data point of the above sensor measurements (dashed horizontal lines). Abrupt increases in SSC, EC, pH and $CH_{4(aq)}$ correspond to outburst events (shaded sections) and reflect sudden drainage of sub-ice-sheet waters and sediments driven by supraglacial meltwater entering the subglacial system (Supplementary Information 2b). The left and right vertical axes correspond to the black and orange datasets, respectively.

outburst events (Supplementary Information 2b; Fig. 1). These outburst events were characterized by pulses in suspended sediment concentrations, electrical conductivity and pH (Fig. 1), indicative of subglacial origin, as previously inferred[18]. The high concentrations of $CH_{4(aq)}$ observed during these events suggest the evacuation of methane-rich subglacial waters from progressively inland sources (Supplementary Information 2b). We attribute the overall decreasing trend in methane concentration following the second outburst event to dilution by rising supraglacial ice-melt inputs to the subglacial system over the melt season. The sustained methane load observed during this period, however, indicates that subglacial methane reserves are not exhausted, despite increases in meltwater discharge (Fig. 1).

The cumulative lateral flux of $CH_{4(aq)}$ from the LG amounted to about 1.87 t (1.64–2.10 t) over the measurement period (reported ranges reflect errors in the measured concentrations and discharge; see Methods). However, we estimate that at least 2.78 t (2.43–3.12 t)—but more probably about 6.28 t (5.19–7.36 t)—of $CH_{4(aq)}$ were laterally transported at the measuring site over the entire 2015 melt season (Fig. 2; see Methods for details). Methane measurements provide conservative estimates of total methane production across the glacier, because recorded concentrations would have been influenced by oxidative and diffusive processes upstream of the measuring site and hence subglacial methane production beneath the catchment is probably larger. On the basis of previously measured microbial oxidation rates[5] and a sustained-flux scenario, we estimate that the bacterial methane sink at the LG amounted to about 1.22 t before subglacial discharge reached the ice margin, which is about 16% of the total methane export at the measuring site over the melt season (Fig. 2; Supplementary Information 2c).

We use scaling relationships between gas transfer velocities and river hydrology[19] to derive conservative approximations of diffusive fluxes of methane from the LG proglacial river. We infer that there is some evasion of methane from subglacial runoff to air spaces in subglacial channels close to the margin[20] and to the atmosphere after emergence at the glacier subglacial portal. We estimate that such atmospheric evasion constitutes the main sink of $CH_{4(aq)}$ compared to microbial oxidation, with diffusive fluxes responsible for at least 1.72 t (0.51–3.19 t) of CH_4 released to the atmosphere between the ice margin and the measuring site (Fig. 2; compared to about 0.09 t CH_4 oxidized for the same distance, or about 1% of exports; data not shown). Recent work on white-water streams has indicated that these traditionally used scaling relationships can greatly underestimate (by several orders of magnitude) diffusive fluxes in white-water systems[21,22]. Considering the high degree of turbulence observed in the LG river (Extended Data Fig. 1), we therefore stress that our estimates here constitute lower-limit values. What is clear is that the LG catchment is a source of atmospheric methane, with our minimum estimates indicating that over 18% (7.5%–26%) of exported methane reaches the atmosphere within 2 km of the ice-sheet margin.

Methane concentrations at the LG fall within the global range reported for streams and rivers (Fig. 3). A recent survey of riverine methane indeed revealed that streams have previously been overlooked as net contributors of atmospheric methane, and they are estimated to emit over 27 Tg CH_4 annually—about 15% and 40% of global wetland and lake effluxes, respectively[6]. The results presented here suggest that streams draining subglacial basins are probably no exception, with the estimated diffusive fluxes of methane at the LG falling in the higher range of reported world averages for rivers, comparable to the large fluxes observed in the Congo basin (Fig. 3, Extended Data Table 1). Because of the high uncertainties surrounding LG methane diffusive fluxes, it is difficult to accurately determine the overall contribution of methane to the atmosphere from the LG catchment and by extension from the GrIS margin as a whole.

To directly compare methane fluxes at the LG with those of other systems, we calculate the catchment-wide areal yield of $CH_{4(aq)}$ that contributed to the observed $CH_{4(aq)}$ lateral flux. When comparing catchment-area-normalized yields of $CH_{4(aq)}$, the lateral $CH_{4(aq)}$ flux from the LG translates into a yield higher than, or within the range of, other large rivers worldwide and highlights that the GrIS may act as a relatively important source of atmospheric methane (Extended Data Table 1; Supplementary Information 1c). Ultimately, the atmospheric footprint of GrIS CH_4 will partly depend on the overall surface area of the ice sheet that contributes to the overall diffusive fluxes, as well as on the magnitude of such fluxes at points of first contact between the atmosphere and subglacial runoff (for example, within open channels beneath the ice).

Stable-isotope analyses ($\delta^{13}C$ and $\delta^{2}H$) reveals that LG methane was microbial in origin, with most samples falling in a well defined range characteristic of acetoclastic methanogenesis, although with some degree of mixing with methane probably produced by a CO_2-reduction pathway (Fig. 4). This mixed origin of methane from CO_2 reduction and acetate fermentation is also supported by molecular evidence from the LG proglacial stream, which identifies the presence of 16S rRNA gene sequences related to both hydrogenotrophic and acetoclastic methanogens (Extended Data Fig. 4; Supplementary Information 2d). A mixed methane source at the LG suggests the availability of several methanogenic substrates beneath the ice, probably derived from the recycling of overridden old carbon (for example, acetate), such as that seen in GrIS marginal lakes[23], potentially supplemented by H_2 gas generated from rock comminution, which has been hypothesized to fuel methanogens beneath ice masses over extended glaciation[24] (see Supplementary Information 2d).

© 2019 Springer Nature Limited. All rights reserved.

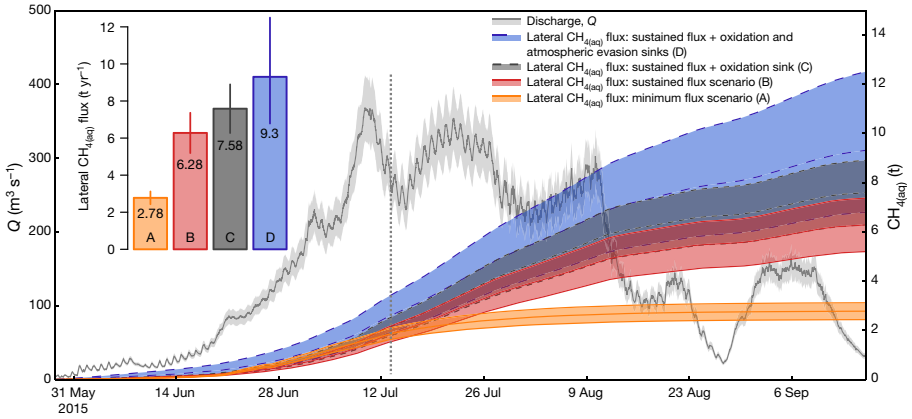


Fig. 2 | Cumulative lateral export of LG $CH_{4(aq)}$ over the 2015 melt season. Orange and red lines correspond to the minimum and sustained methane flux scenarios, respectively, at the measuring site (see Methods). Dark-grey lines represent a scenario that accounts for a methanotrophic methane sink on a sustained-flux scenario and represent the expected lateral methane flux that would have occurred without a methanotrophic sink. Blue lines correspond to a scenario that accounts for the combined estimated methanotrophic and diffusive flux sinks of methane before reaching the measuring site, added to a sustained-flux scenario. The vertical dotted line marks the last day of $CH_{4(aq)}$ sensor measurements (13 July). The width of the shaded areas corresponds to errors from sensor measurements and estimates of gas transfer velocities (see Methods). The pale-grey time series denotes discharge measurements over the entire melt season. The annual methane fluxes depicted in the bar plot correspond to the cumulative fluxes at the end of the melt season for each of the estimated scenarios; error bars correspond to the range depicted by the shaded areas.

Partial oxidation during transit from the subglacial system probably enriched the sampled methane with heavier stable isotopes[25] (Supplementary Information 2c), yet there is no strong isotopic trend that conclusively identifies methanotrophy as a major control on the isotopic signatures observed here (Fig. 4, Extended Data Fig. 3). This contrasts with patterns that we observed for stagnant waters beneath the LG proglacial river ice (Extended Data Fig. 3) and waters sampled from Antarctic Subglacial Lake Whillans (Supplementary Information 2c). We infer the limited methanotrophic signature observed here to reflect the largely anoxic conditions at the sites of methane production (and thus limited aerobic oxidation of methane) and the rapid evacuation of methane from the production site via a fast and efficient drainage system (Supplementary Information 2b).

The impact of subglacial methane on atmospheric concentrations partially depends on the presence of methane hydrates beneath ice sheets, as catastrophic methane hydrate destabilization during periods of rapid ice thinning could result in very large fluxes of methane to the atmosphere[3,4]. We used a one-dimensional reaction-transport model to identify the conditions required to allow methane hydrate formation beneath the LG catchment. Our results indicate that relatively high methanogenic rates (larger than those observed in Greenland basal ice incubation experiments[17]; Extended Data Fig. 5) and thick sediment layers (at least several tens of metres) are required to produce and sustain methane hydrates beneath the LG catchment (Supplementary Information 2e, f). The high methane flux that would be generated at the ice–sediment interface under methane hydrate conditions (estimated to be 10 to 1,000 times larger than the observed lateral flux, depending on hydrate conditions; Extended Data Fig. 6) makes it unlikely that a substantial portion (if any) of the exported CH_4 measured from the LG comes from subglacial methane hydrates. Importantly, however, the model results suggest that conditions favourable to hydrate formation are probably present in other regions of the GrIS with sustained thick ice cover (for example, for more than 10,000 years) and with thick sedimentary layers (for example, ref. [27]; Supplementary Information 2f).

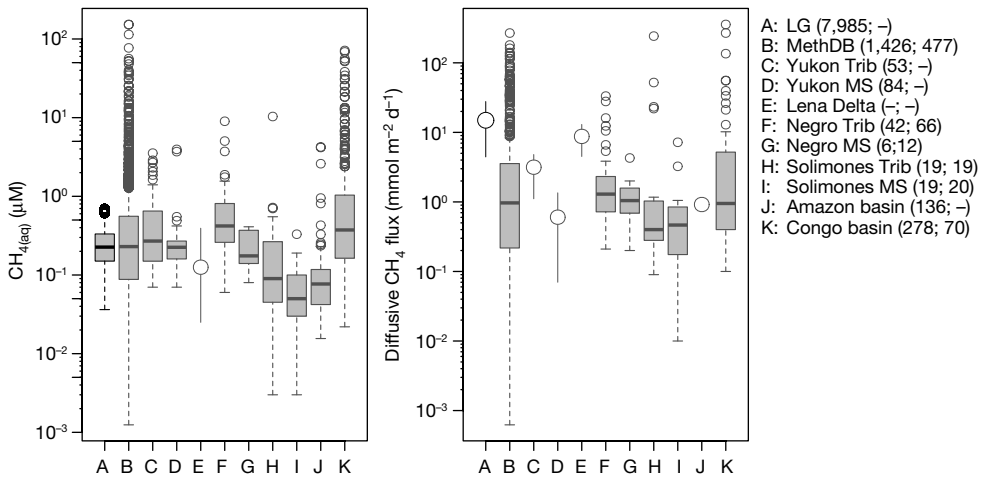


Fig. 3 | Box plots of $CH_{4(aq)}$ concentrations and diffusive fluxes for the LG and other major world river systems. The box mid-lines represent medians; the interquartile range (IQR) is represented by the lower and upper box boundaries, which denote the 25th and 75th percentiles, respectively; whiskers indicate confidence intervals 1.5 times the IQR, and points are outliers. The parentheses next to the names of the rivers give the number of observations for concentrations (left) and fluxes (right). Where no raw data are available, averages and reported ranges are depicted by circles and error bars (see Supplementary Information 1c for details). 'MethDB' refers to a worldwide $CH_{4(aq)}$ dataset for rivers[6]; 'Trib' and 'MS' refer to the tributaries and mainstems of the rivers, respectively.

© 2019 Springer Nature Limited. All rights reserved.

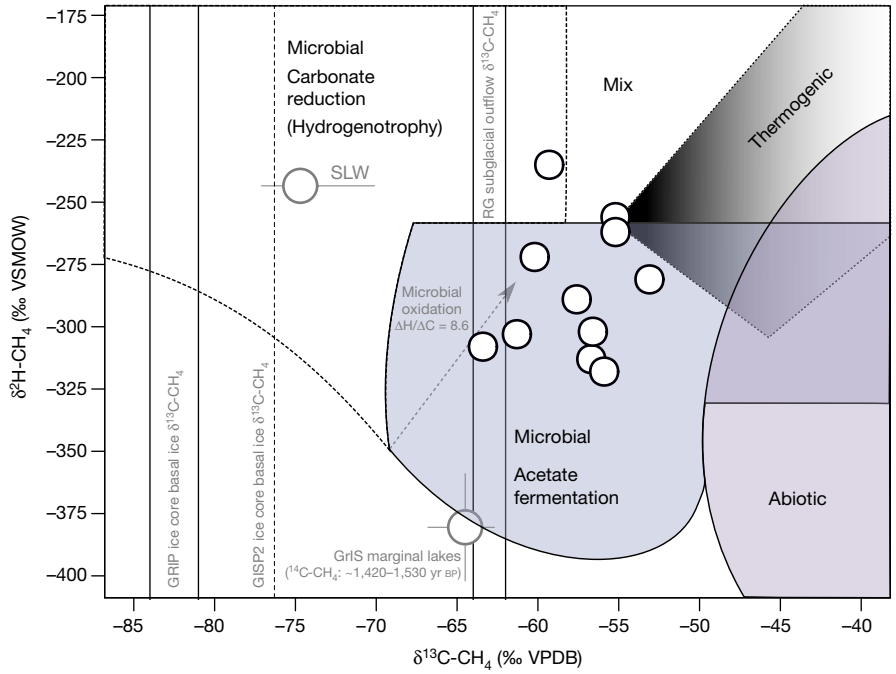


Fig. 4 | Carbon–hydrogen isotopic diagram of LG $CH_{4(aq)}$. Black-border points denote dual stable-isotope values ($\delta^{13}C$ and $\delta^{2}H$) for LG $CH_{4(aq)}$ samples (sample values are summarized in Extended Data Table 2). Average $\delta^{13}C$-CH_4 and $\delta^{2}H$-CH_4 values and ranges from Subglacial Lake Whillans (SLW) in Antarctica[3] and GrIS marginal lakes[24] are added as references (grey-border points), as well as $\delta^{13}C$-CH_4 data from GrIS ice-core basal ice[12,23] and from the subglacial outflows of the Greenland Russell Glacier (RG)[6] (marked by vertical lines). The estimated carbon age (years before present, BP) from ^{14}C analyses of methane from GrIS marginal lakes[24] is indicated next to point. The arrow denotes the microbial oxidation effect on CH_4 stable-isotope signatures; $\Delta H/\Delta C$ is the gradient (change in $\delta^{2}H$-CH_4 over change in $\delta^{13}C$-CH_4) of the arrow[26]. The classification zones and definitions of methane origins are derived and adapted from refs [25,26]. VSMOW, Vienna standard mean ocean water; VPDB, Vienna Pee Dee belemnite.

Using high-resolution in situ sensor measurements, we show that an extensive area of the GrIS continuously releases methane-supersaturated runoff from its bed during the melt season. Our results constitute the first measurements of sustained methane export from an ice-sheet catchment and highlight the need to better gauge the footprint of ice sheets on current methane budgets. The release of several tonnes of microbial methane from beneath the GrIS represents one of the strongest lines of evidence to date for considerable microbial production of methane in subglacial ecosystems and reinforces the view that large methane reserves may accumulate beneath past and present-day ice sheets[3,7]. This methane can reach the atmosphere where fast-flowing drainage networks enable its rapid transport beyond the ice margin before being oxidized to carbon dioxide, either driven by supraglacial forcing in the GrIS ablation zone or potentially also during episodic subglacial lake drainage events in Antarctica[28]. The influence of meltwater discharge on methane export observed here further suggests that projected increases in warming and melting rates could also lead to increases in subglacial methane release to the atmosphere. Our finding that subglacial environments in Greenland can generate high levels of methane emphasizes the need to directly measure methane reserves in subglacial systems containing high quantities of organic carbon, such as the thick sedimentary basins beneath the Antarctic Ice Sheet, where much larger amounts of methane, as hydrates, are expected to be present[4].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0800-0.

Received: 23 October 2017; Accepted: 8 November 2018;
Published online 2 January 2019.

1. Kirschke, S. et al. Three decades of global methane sources and sinks. *Nat. Geosci.* **6**, 813–823 (2013).
2. Schaefer, H. et al. A 21st century shift from fossil-fuel to biogenic methane emissions indicated by $^{13}CH_4$. *Science* **352**, 80–84 (2016).
3. Wadham, J. L., Tranter, M., Tulaczyk, S. & Sharp, M. Subglacial methanogenesis: a potential climatic amplifier? *Global Biogeochem. Cy.* **22**, GB2021 (2008).
4. Wadham, J. L. et al. Potential methane reservoirs beneath Antarctica. *Nature* **488**, 633–637 (2012).
5. Dieser, M. et al. Molecular and biogeochemical evidence for methane cycling beneath the western margin of the Greenland Ice Sheet. *ISME J.* **8**, 2305–2316 (2014).
6. Stanley, E. H. et al. The ecology of methane in streams and rivers: patterns, controls, and global significance. *Ecol. Monogr.* **86**, 146–171 (2016).
7. Weitemeyer, K. A. & Buffett, B. A. Accumulation and release of methane from clathrates below the Laurentide and Cordilleran ice sheets. *Global Planet. Change* **53**, 176–187 (2006).
8. Michaud, A. B. et al. Microbial oxidation as a methane sink beneath the West Antarctic Ice Sheet. *Nat. Geosci.* **10**, 582–586 (2017).
9. Petrenko, V. V. et al. Minimal geological methane emissions during the Younger Dryas–Preboreal abrupt warming event. *Nature* **548**, 443–446 (2017).
10. Portnov, A., Vadakkepuliyambatta, S., Mienert, J. & Hubbard, A. Ice-sheet-driven methane storage and release in the Arctic. *Nat. Commun.* **7**, (2016).
11. Souchez, R., Lemmens, M. & Chappellaz, J. Flow-induced mixing in the GRIP basal ice deduced from the CO_2 and CH_4 records. *Geophys. Res. Lett.* **22**, 41–44 (1995).
12. Miteva, V., Teacher, C., Sowers, T. & Brenchley, J. Comparison of the microbial diversity at different depths of the GISP2 Greenland ice core in relationship to deposition climates. *Environ. Microbiol.* **11**, 640–656 (2009).
13. Christner, B. C., Montross, G. G. & Priscu, J. C. Dissolved gases in frozen basal water from the NGRIP borehole: implications for biogeochemical processes beneath the Greenland Ice Sheet. *Polar Biol.* **35**, 1735–1741 (2012).
14. Schmidt, M., Linke, P. & Esser, D. Recent development in IR sensor technology for monitoring subsea methane discharge. *Mar. Technol. Soc. J.* **47**, 27–36 (2013).
15. Cowton, T., Nienow, P., Bartholomew, I., Sole, A. & Mair, D. Rapid erosion beneath the Greenland ice sheet. *Geology* **40**, 343–346 (2012).
16. Kohler, T. J. et al. Carbon dating reveals a seasonal progression in the source of particulate organic carbon exported from the Greenland Ice Sheet. *Geophys. Res. Lett.* **44**, 6209–6217 (2017).
17. Stibal, M. et al. Methanogenic potential of Arctic and Antarctic subglacial environments with contrasting organic carbon sources. *Glob. Change Biol.* **18**, 3332–3345 (2012).
18. Bartholomew, I. et al. Supraglacial forcing of subglacial drainage in the ablation zone of the Greenland ice sheet. *Geophys. Res. Lett.* **38**, L08502 (2011).
19. Raymond, P. A. et al. Global carbon dioxide emissions from inland waters. *Nature* **503**, 355–359 (2013); erratum **507**, 387 (2013).

© 2019 Springer Nature Limited. All rights reserved.

20. Chandler, D. M. et al. Evolution of the subglacial drainage system beneath the Greenland Ice Sheet revealed by tracers. *Nat. Geosci.* **6**, 195–198 (2013).
21. Hall, R. O., Kennedy, T. A. & Rosi-Marshall, E. J. Air–water oxygen exchange in a large whitewater river. *Limnol. Oceanogr. Fluids Environ.* **2**, 1–11 (2012).
22. Maurice, L., Rawlins, B. G., Farr, G., Bell, R. & Gooddy, D. C. The influence of flow and bed slope on gas transfer in steep streams and their implications for evasion of CO_2. *J. Geophys. Res. Biogeosci.* **122**, 2862–2875 (2017).
23. Walter Anthony, K. M., Anthony, P., Grosse, G. & Chanton, J. Geologic methane seeps along boundaries of Arctic permafrost thaw and melting glaciers. *Nat. Geosci.* **5**, 419–426 (2012).
24. Telling, J. et al. Rock comminution as a source of hydrogen for subglacial ecosystems. *Nat. Geosci.* **8**, 851–855 (2015); erratum 8, 981 (2015).
25. Whiticar, M. J. Carbon and hydrogen isotope systematics of bacterial formation and oxidation of methane. *Chem. Geol.* **161**, 291–314 (1999).
26. Etiope, G. & Sherwood Lollar, B. Abiotic methane on Earth. *Rev. Geophys.* **51**, 276–299 (2013).
27. Walter, F., Chaput, J. & Lüthi, M. P. Thick sediments beneath Greenland's ablation zone and their potential role in future ice sheet dynamics. *Geology* **42**, 487–490 (2014).
28. Wingham, D. J., Siegert, M. J., Shepherd, A. & Muir, A. S. Rapid discharge connects Antarctic subglacial lakes. *Nature* **440**, 1033–1036 (2006).

Acknowledgements We thank all of those who assisted with fieldwork at LG, especially J. Hatton, as well as F. Sgouridis and J. Williams at the LOWTEX laboratory of the University of Bristol. This research is part of the UK NERC-funded DELVE programme (NERC grant NE/I008845/1 to J.L.W.). G.L.-G. was funded by the University of Bristol Scholarship Programme and a FRQNT Scholarship (number 185136). The work was also supported by a Leverhulme research fellowship to J.L.W., a UK NERC grant (NE/J02399X/1) to A.M.A. for DNA analyses, as well as Czech Science Foundation grants (GACR; 15-17346Y and 18-12630S) to M.S. Isotopic analyses were conducted by G. Lacrampe-Couloume at the University of Toronto with support provided by the Natural Sciences and Engineering Research Council of Canada (NSERC) to B.S.L. We also thank the Kangerlussuaq International Science Station, especially R. Møller, for support with field logistics, as well as M. A. Cooper, M. Macdonald and S. Hoffer for comments.

Reviewer information *Nature* thanks J. Crawford and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions J.L.W. and G.L.-G. designed the study. B.S.L. supervised the stable-isotope analyses. S.A. performed the reaction-transport hydrate model calculations. P.F. assisted in the interpretation and analysis of the CONTROS HydroC CH_4 raw results. G.L.-G., J.T., E.A.B., A.D.B., M.C.M. and J.R.H. conducted field logistical preparations. J.L.W., J.T. and M.S. led the 2015 Greenland field campaign. G.L.-G., A.D.B., A.J.T., J.T., E.A.B., J.R.H., T.J.K., J.D.Z. and M.S. collected the sensor field data. G.L.-G. and J.D.Z. collected manual water samples in the field. G.L.-G. and A.M.A. analysed molecular data. G.L.-G. performed the data analysis and wrote the manuscript with significant contribution from all co-authors.

Competing interests P.F. works for the sensor manufacturer Kongsberg Maritime Contros, but the sensor data discussed here were validated by independent measurements. The other authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0800-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0800-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to G.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

Site description and hydrogeochemical analyses. The hydrology of the LG has been extensively studied and described previously (see Supplementary Information 1a). A detailed description of the proglacial study site, as well as the hydrological and geochemical monitoring performed during the 2015 melt season, can be found in two parallel studies[16,29]. In brief, a suite of hydrogeochemical sensors recording pH (Honeywell Durafet), water temperature (Aanderaa and Campbell Scientific), electrical conductivity (Campbell Scientific 547) and turbidity (Partech C) were deployed in the LG proglacial river, about 1.6 km downstream from the subglacial ice portal at the glacier's terminus (Extended Data Fig. 1). Turbidity measurements were converted to suspended sediment concentrations by calibration against manual sediment samples collected over the span of the melting season, as in ref. [30]. Discharge measurements were derived from pressure transducers (Druck and Hobo) and stage sensors (Campbell Scientific SR50A) fixed in a bedrock section about 2 km downstream from the glacier's terminus. Stage measurements were converted to discharge values using a stage–discharge rating curve generated from calibration against repeated rhodamine dye injections over the full range of river stages during the melt season, as in ref. [18]. Uncertainties (root-mean-square deviation) on discharge measurements were calculated to be about 12.1%.

Manual sampling. Manual samples were collected a few metres (about 5–10 m) upstream of the HydroC sensor. Water samples were collected inside pre-evacuated (at most 500 mTorr) 120-ml borosilicate vials sealed with 2-cm-thick butyl-rubber stoppers, pre-flushed with 5.0 grade argon and pre-poisoned with about 24 mg of $HgCl_2$ to fix the samples and prevent any microbial activity from affecting the gases post-sampling, after the method of ref. [31]. 10 ml (at room temperature and pressure) of helium (grade 5.0) was added to the evacuated vials to maintain a headspace during sampling. Most water samples ($n = 53$) were collected using a peristaltic pump (Portapump-810, Williamson Manufacturing) equipped with silicone tubing; a small number of samples were collected using plastic syringes ($n = 2$) or passively, using the vials' vacuum pressure by directly piercing the septum of submerged vials with a needle ($n = 8$). Vials containing apparent air contamination or vacuum loss (for example, resulting in abnormally large headspace post-sampling) were excluded from analyses. Samples for stable-isotope analysis were collected as above ($n = 9$ collected using the peristaltic pump, $n = 2$ using syringes).

Methane concentrations were calculated using the headspace method. Headspace samples were analysed on an Agilent 7980A gas chromatograph equipped with a Porapak Q 80–100 mesh, 2.5 m × 2.0 mm stainless-steel column and flame-ionization detector. Standard curves were calculated from certified (±5%) gas-standard measurements. Gas concentrations were converted to molar concentrations using the ideal gas law, and dissolved methane concentrations were obtained using Bunsen coefficients[32]. Internal vial pressures were calculated using the ideal gas law from the difference between the headspace volume post-sampling and the theoretical headspace volume of 10 ml at 1 atm and 20 °C. An average internal pressure of 3.5 ± 0.9 atm (standard deviation) was assigned to all manual samples for calculations.

CONTROS HydroC CH_4 sensor. Methane measurements were performed using a CONTROS HydroC CH_4 system (Kongsberg Maritime), an optical (infrared), headspace-based underwater sensor. An underwater pump (SBE 5T, Sea-Bird Scientific) mounted to the sensor continuously feeds water to the membrane equilibrator. Dissolved gases diffuse through a composite membrane into the internal gas circuit, where partial pressure is measured using tunable diode laser absorption technology[14]. The CONTROS HydroC sensor was deployed completely submerged, within a solid metallic cage moored by cables attached to boulders on the river bank, with the sensor head facing the river current (Extended Data Fig. 1). Measurements were logged every minute between 19 May and 4 June; the logging interval was changed to 5 min on 4 June until the end of the measuring period, on 13 July.

The ideal gas law and Bunsen coefficients were used to convert microatmosphere-scale measurements (Extended Data Fig. 7c) to molar concentrations (Fig. 1). Water temperatures (±0.05 °C reported accuracy) were recorded using an Aanderaa Optode 3830 sensor deployed in parallel (Extended Data Fig. 7a). The reported overall uncertainty of CONTROS HydroC CH_4 is 2 μatm (about 5 nM) or ±3% of the reading, whichever is greater.

Calculation of lateral methane flux. The $CH_{4(aq)}$ measurements stopped on 13 July. $CH_{4(aq)}$ fluxes estimated during the rest of the ablation season were based on two scenarios: (i) assuming that methane levels would immediately decrease until they reached river baseline concentrations on 15 September (last discharge measurement), or (ii) assuming that methane levels would continue to follow a discharge-dependent trend for the duration of the ablation season.

(i) Constant concentration decrease (annual lateral CH_4 flux of 2.78 t). In the constant-concentration-decrease scenario, a baseline $CH_{4(aq)}$ concentration was set on the basis of manual water samples collected during a return visit to the sampling site on 28 October, when the proglacial river was partially frozen and no runoff contribution to the proglacial river stream was apparent. The October concentrations averaged about 18.5 nM (beneath river ice at that time; $n = 6$).

The minimum flux scenario was calculated using a natural-logarithm decrease of the form

$$y = Ce^{-kt}$$

where y is the methane flux (for example, in grams per second), C is the last flux measurement, obtained on 13 July (that is, 0.71 g s^{-1}), t is the time elapsed between 13 July and the measurement of flux y, and k is the reaction constant, obtained assuming a baseline concentration of 18.5 nM and using a discharge of 32 m^3 s^{-1} (last discharge measurement, on 15 September).

(ii) Sustained flux (annual lateral CH_4 flux of 6.28 t). The sustained-flux scenario was calculated using the discharge-weighted mean $CH_{4(aq)}$ concentration of 271 ± 34 nM, obtained from measurements up to 13 July; the error reflects errors on discharge measurements (12.1%) as well as measurement errors of the HydroC CH_4 sensor (2 μatm or 3%, whichever is greater).

Estimation of methane sink via methanotrophic oxidation. The methane concentrations recorded at the LG probably underestimate the original methane levels present beneath the catchment because of the water travel time between the subglacial methane source and the measurement site. In addition to atmospheric evasion of methane, aerobic microbial oxidation of methane would have lowered methane concentrations before reaching the observation site, once fully oxygenated meltwater runoff entered the subglacial system (O_2 concentrations in runoff were either near atmospheric equilibrium or supersaturated for most of the monitoring period; Extended Data Fig. 7a). Methanotrophy was observed qualitatively in a small number of unfixed river samples collected in parallel with fixed manual samples. Analyses at the home laboratory showed $CH_{4(aq)}$ concentrations decreased by a factor of up to 100 in unfixed samples compared with the fixed vials (data not shown). However, no time series incubation was set up and consequently no methanogenic rates were calculated for the LG site.

The quantity of methane oxidized by methanotrophic bacteria before reaching the measuring site was estimated using the methanotrophic rate reported for the marginal stream of the neighbouring Russell Glacier (that is, 0.32 μM d^{-1})[5]. Justifications for using the Russell Glacier oxidation rate are discussed in Supplementary Information 2c. The time during which runoff was subject to methane oxidation (that is, water travel time) was estimated from water velocities and subglacial drainage evolution calculated on the basis of a previous study at the LG[20]. We assumed that subglacial aerobic methane oxidation occurs between the location of supraglacial runoff–input, where oxygenated supraglacial waters enter the subglacial system, and the measuring site located 1.6 km downstream of the LG glacier terminus.

Water velocities were calculated using the relationship between maximum tracer velocity (v_{05}) and cumulative discharge ($\sum Q$), described for the gaseous tracer SF_6 in ref. [20], which takes the form

$$v_{05} = A \ln(\Sigma Q) + B$$

with the regression parameters A and B calculated to be 0.235 m s^{-1} and −3.59 m s^{-1}, respectively[20]. We fixed the minimum velocity at 0.4 m s^{-1}, which corresponds to the minimum v_{05} calculated in ref. [20] for tracer injections performed 7 km inland from the LG portal at times of low cumulative discharge.

We estimated the inland evolution of an efficient channelized subglacial hydrological system on the basis of the relationship between cumulative discharge and v_{05} at moulin injection sites (see figure 2a in ref. [20]). We derived the progression of supraglacial water inputs using the lowest value of cumulative discharge, observed where v_{05} at an injection site fell onto the regression line of v_{05} on the cumulative discharge for the L7 injection in ref. [20] (see supplementary figure 2.8 in ref. [20]). That is, we assumed that the channelized subglacial channel would reach 7 km at a cumulative discharge of 1.9×10^7 m^3, 14 km at 9.4×10^7 m^3 and 41 km at 7.8×10^8 m^3 on the basis of supplementary figure 2.8 and supplementary table 1 in ref. [20]. Although we acknowledge that such calculations are approximate at best, they allow the use of a dynamic distance of travel during the melt season. We fixed the maximum travel distance at 41 km from the LG terminus, after which the LG subglacial system is considered to become primarily inefficient and distributed for the duration of the ablation season[20]. To account for potential methane sources and methanotrophic activity occurring downstream of the supraglacial-runoff input into the subglacial channelized system, we used an average distance of travel in our calculation (that is, half of the distance of travel obtained from the cumulative-discharge calculations above).

Calculation of diffusive methane flux. Accurately calculating methane losses due to atmospheric evasion was beyond the scope of the present study, and therefore flux numbers should be considered conservative estimates of the amount of methane originally generated and exported from the LG catchment.

© 2019 Springer Nature Limited. All rights reserved.

Diffusive fluxes for the LG stream were estimated following the approach in ref. [19], which estimates gas transfer velocity coefficients (k) from the stream slope and water velocity (fitted equation (5) in ref. [33]). Fluxes were estimated for the first 1.6 km of the proglacial river, from the ice margin to the measuring site. The stream slope was obtained from Google Earth and approximated as 0.04; slope values of 0.01, 0.03 and 0.05 were used to generate minimum, medium and maximum k values, respectively. A water velocity of 1 m s^{-1} was used, which corresponds to the discharge weighted mean of subglacial water velocities (v_{05}) used for methanotrophic sink calculations (see above).

Methane gas-transfer velocities (k_{CH4}) were converted from the calculated k_{600} values following relationships between Schmidt numbers and k for CO_2 and CH_4 (see equations (2) and (3) in ref. [33]); Schmidt numbers were calculated using an average water temperature of 0.22 °C (Extended Data Fig. 7d)[34]. Minimum, medium and maximum slope values, as well as standard deviations on k_{600} equation parameters[33], resulted in minimum, medium and maximum k_{CH4} of 16 m d^{-1}, 49 m d^{-1} and 84 m d^{-1}, respectively.

Methane diffusive fluxes were calculated using the discharge-weighted mean $CH_{4(aq)}$ concentration for the observation period (271 nM) and assuming an atmospheric methane concentration of 1.8 p.p.m. by volume (resulting in an equilibrium concentration of about 4.6 nM). Diffusive flux upstream of the measuring site was calculated using 1-m retroactive bins, adjusting upstream dissolved methane concentrations for methane loss by both diffusive flux and microbial oxidation losses in downstream bins; a fixed river width of 40 m, water velocity of 1 m s^{-1} and average discharge of 150 m^3 s^{-1} were used in the calculations. The reported diffusive flux values correspond to the average flux calculated for the 1.6 km of stream for each minimum, medium and maximum scenarios. Cumulative fluxes were calculated for the discharge measurement period (that is, about 110.5 days) and normalized to estimated water velocities (see previous section). Details on the diffusive fluxes of other world rivers can be found in Supplementary Information 1c.

Stable-isotope analyses. Analyses for $\delta^{13}C$ values were performed by continuous-flow compound-specific carbon-isotope-ratio mass spectrometry with a Finnigan MAT 252 mass spectrometer interfaced with a Varian 3400 capillary gas chromatograph. Hydrocarbons were separated by a Poraplot Q column (25 m × 0.32 mm internal diameter) with the following temperature programme: initial temperature 40 °C, hold for 1 min, increase to 190 °C at 5 °C min^{-1}, hold for 5 min. The total error, incorporating both accuracy and reproducibility, is ±0.5‰ with respect to the Vienna Pee Dee belemnite standard[35]. The δ^2H analysis was performed on a continuous-flow compound-specific hydrogen-isotope mass spectrometer that consists of an HP 6890 gas chromatograph interfaced with a micropyrolysis furnace (1,465 °C) in line with a Finnigan MAT Delta$^+$-XL isotope-ratio mass spectrometer. H_2 and CH_4 were separated by a molecular sieve 5A column (25 m × 0.32 mm internal diameter) with a carrier gas flow rate of 1.2 ml min^{-1} and the temperature programme: initial temperature 20 °C, hold for 5 min, followed by an increase to 280 °C at 25 °C min^{-1}. Higher hydrocarbons were separated using the same column and temperature programme as those used in the carbon isotope analysis. The total error, incorporating both accuracy and reproducibility, for the hydrogen isotope analysis is ±5‰ with respect to V-SMOW[31].

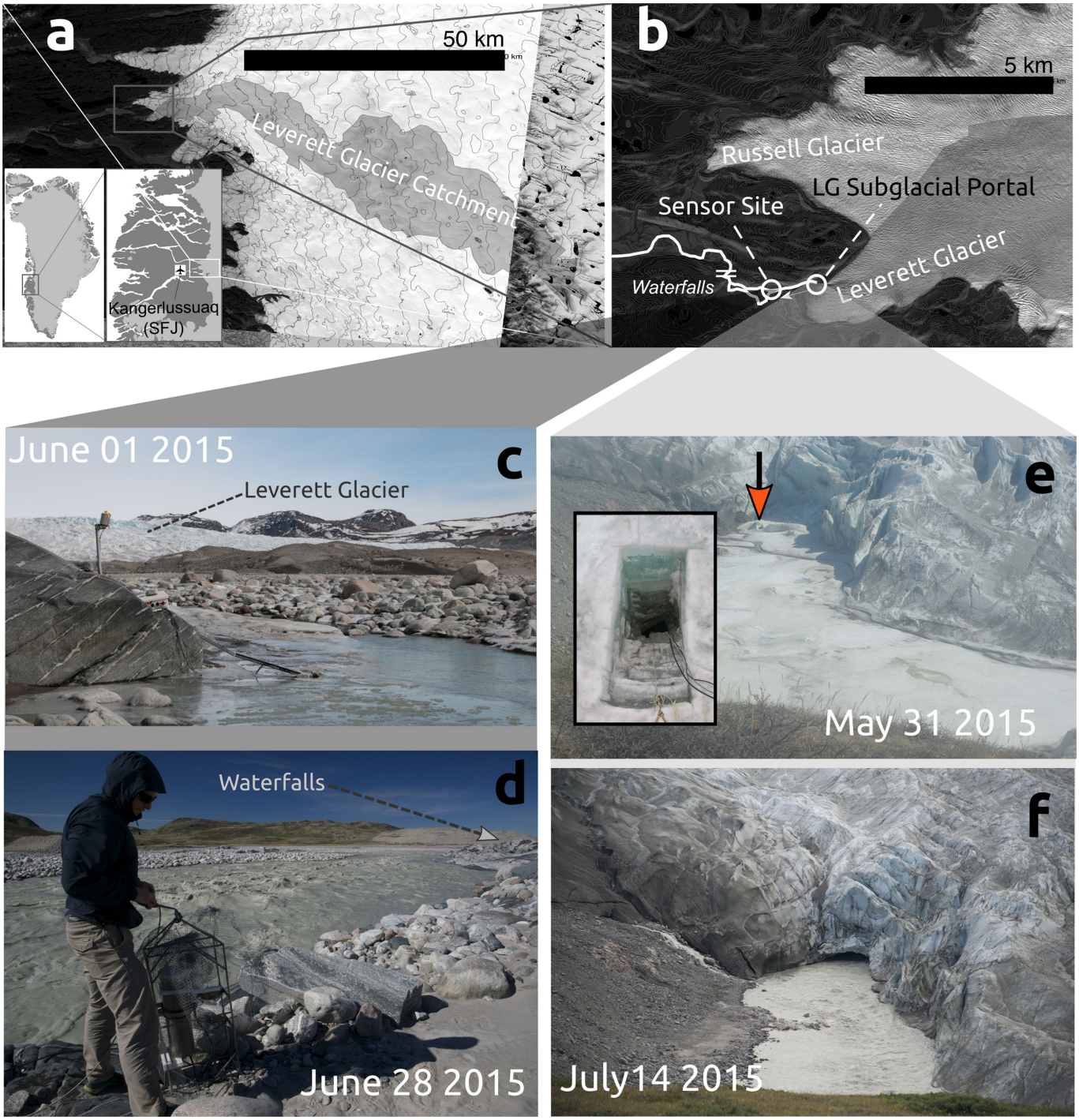
Methane hydrates. To evaluate the potential for hydrate formation beneath the LG catchment, we used a one-dimensional reaction-transport model that was originally developed for simulating hydrate formation in marine sediments[36] and has previously been adapted for subglacial Antarctica[4]. We assumed physical properties for sediments similar to those previously used for ocean sediment modelling[36]. Extended Data Table 3 summarizes site-specific model parameters, their model values and units. The model solves the one-dimensional diffusion–advection–reaction equations for dissolved methane, gaseous methane and methane hydrates. The implemented reaction network accounts for a constant methane production rate R_{xn} over a predefined sediment depth z_{xn}, methane hydrate, as well as methane gas formation and dissociation. At the upper boundary, the boundary concentrations were set to zero (that is, Dirichlet boundary condition), reflecting warm-based conditions and allowing for diffusive flux of methane through the ice-sediment interface. In addition, the initial conditions for dissolved and gaseous methane and methane hydrates were set to zero. A 'best case' scenario was designed to reflect optimal, but plausible, physical and biogeochemical conditions for hydrate formation to assess the maximum potential for hydrate accumulation in the catchment. More specifically, we assigned a thick methanogenic sediment layer beneath the catchment (that is, up to 100 m), a 10,000-year ice-sheet overburden to allow for hydrate evolution, complete anoxic conditions, an overlaying ice thickness set to 1,000 m (ice thickness over the LG catchment exceeds 1,000 m at about 39 km from the ice margin[20]), a basal temperature of −1 °C, and absence of a methane sink within the sediment layer (for example, no anaerobic oxidation of methane). This 'best case' model setup was run over a wide range of constant methane production rates (R_{xn} of 10^{-17} to 10^{-13} grams CH_4 per gram of wet sediment per second) to determine the order of magnitude of methane production rates required to accumulate hydrates. After this initial screening, methane production rates were varied systematically between 10^{-15} to 10^{-14} grams CH_4 per gram of wet sediment per second.

Data availability

The data used in this article are available from the corresponding author on request. The 16S rRNA gene sequence data are available from the NCBI Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) under BioProject PRJNA495593 (BioSamples SAMN10228172-SAMN10228185; SAMN10228190-SAMN10228206).

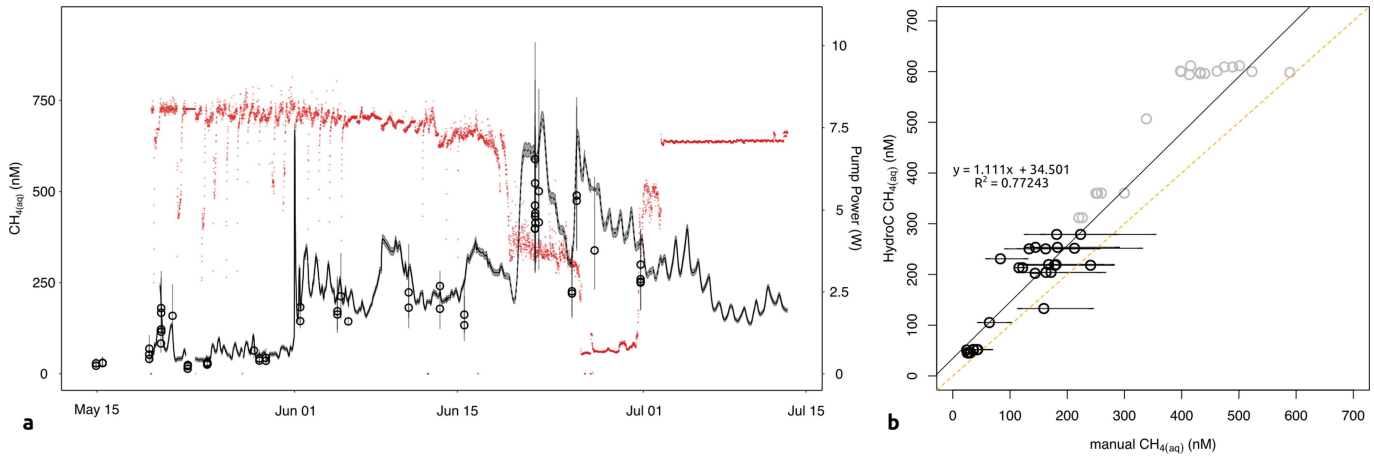
29. Beaton, A. D. et al. High-resolution in situ measurement of nitrate in runoff from the Greenland Ice Sheet. *Environ. Sci. Technol.* **51**, 12518–12527 (2017).
30. Hawkings, J. R. et al. Ice sheets as a significant source of highly reactive nanoparticulate iron to the oceans. *Nat. Commun.* **5**, 3929 (2014).
31. Ward, J. A. et al. Microbial hydrocarbon gases in the Witwatersrand Basin, South Africa: implications for the deep biosphere. *Geochim. Cosmochim. Acta* **68**, 3239–3250 (2004).
32. Wiesenburg, D. A. & Guinasso, N. L., Jr. Equilibrium solubilities of methane, carbon monoxide, and hydrogen in water and sea water. *J. Chem. Eng. Data* **24**, 356–360 (1979).
33. Raymond, P. A. et al. Scaling the gas transfer velocity and hydraulic geometry in streams and small rivers. *Limnol. Oceanogr. Fluids Environ.* **2**, 41–53 (2012).
34. Wanninkhof, R. Relationship between wind speed and gas exchange over the ocean revisited. *Limnol. Oceanogr. Methods* **12**, 351–362 (2014).
35. Sherwood Lollar, B., Hirschorn, S. K., Chartrand, M. M. G. & Lacrampe-Couloume, G. An approach for assessing total instrumental uncertainty in compound-specific carbon isotope analysis: implications for environmental remediation studies. *Anal. Chem.* **79**, 3469–3475 (2007).
36. Davie, M. K. & Buffett, B. A. A numerical model for the formation of gas hydrate below the seafloor. *J. Geophys. Res.* **106**, 497–514 (2001).
37. Tedesco, M. et al. Evidence and analysis of 2012 Greenland records from spaceborne observations, a regional climate model and reanalysis data. *Cryosphere* **7**, 615–630 (2013).
38. Palmer, S., Shepherd, A., Nienow, P. & Joughin, I. Seasonal speedup of the Greenland Ice Sheet linked to routing of surface water. *Earth Planet. Sci. Lett.* **302**, 423–428 (2011).

© 2019 Springer Nature Limited. All rights reserved.



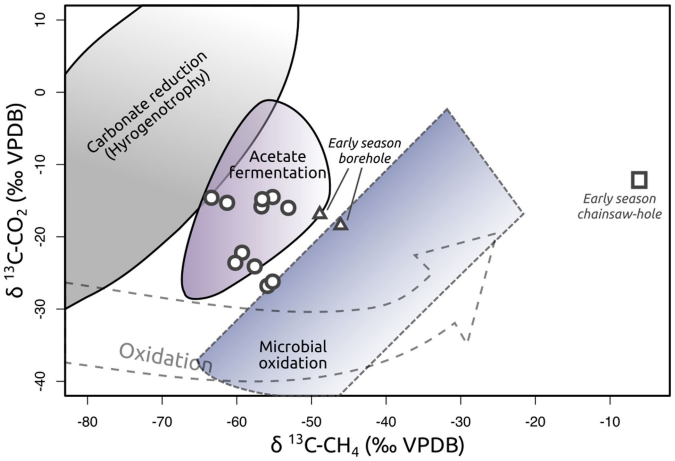
Extended Data Fig. 1 | Leverett Glacier and proglacial stream. **a**, Leverett Glacier (LG), with catchment boundaries[38] outlined in grey; 'SFJ' denotes the Kangerlussuaq airport. **b**, Zoomed image of the LG, with the sampling site and portal marked by dots. **c**, Sensor deployment site during the early melt season, with the LG visible in the background; the image faces upstream. **d**, Sensor deployment site in late June; the image faces downstream. Also visible is the HydroC sensor inside a steel cage, during inspection and before redeployment. **e**, LG portal in late May, while still covered with both glacial and river ice. The photograph was taken one hour before the appearance of the glacial upwelling (see Supplementary Information 2b). The arrow marks the location of the chainsawed hole, shown in the inset (photograph taken on 10 May 2015). **f**, LG portal in mid-July 2015. Map images from USGS/NASA Landsat.

© 2019 Springer Nature Limited. All rights reserved.



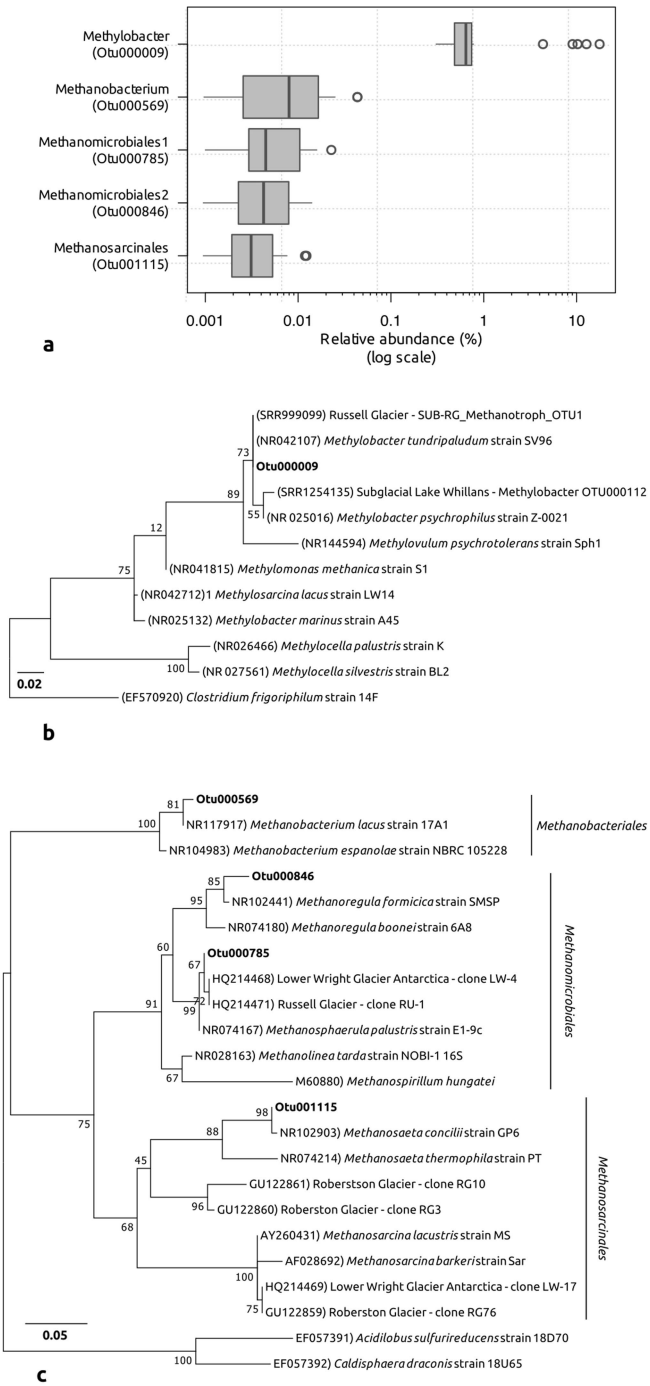
Extended Data Fig. 2 | Comparison of $CH_{4(aq)}$ concentrations measured with the HydroC sensor and from manual samples. a, $CH_{4(aq)}$ time series. Red points correspond to the HydroC pump power during operation. The continuous line depicts HydroC measurements, with the dashed section corresponding to times when the sensor experienced low pump power and thus a reduced water flow induced by the pump (19 June to 1 July). Open circles correspond to manual samples. The thin grey-shaded area around the $CH_{4(aq)}$ time series corresponds to the uncertainty of the HydroC measurements (about 3%). The uncertainty on manual measurements, indicated by the error bars, reflects the error on vial internal pressures and volumes (119 ± 0.76 ml, where the uncertainty represents standard deviation; internal pressures are derived from volumes—see Methods for details). **b**, Regression plot between the HydroC and manual sample measurements. Only manual samples taken during times when the HydroC pump power was above about 7 W were considered for the regression (black circles, black line); grey circles correspond to samples taken during times of lower pump power. Horizontal error bars reflect errors on manual measurements; vertical error bars are smaller than the size of the markers. The orange dashed line depicts a hypothetical 1:1 relationship between the sensor and manual measurements.

© 2019 Springer Nature Limited. All rights reserved.



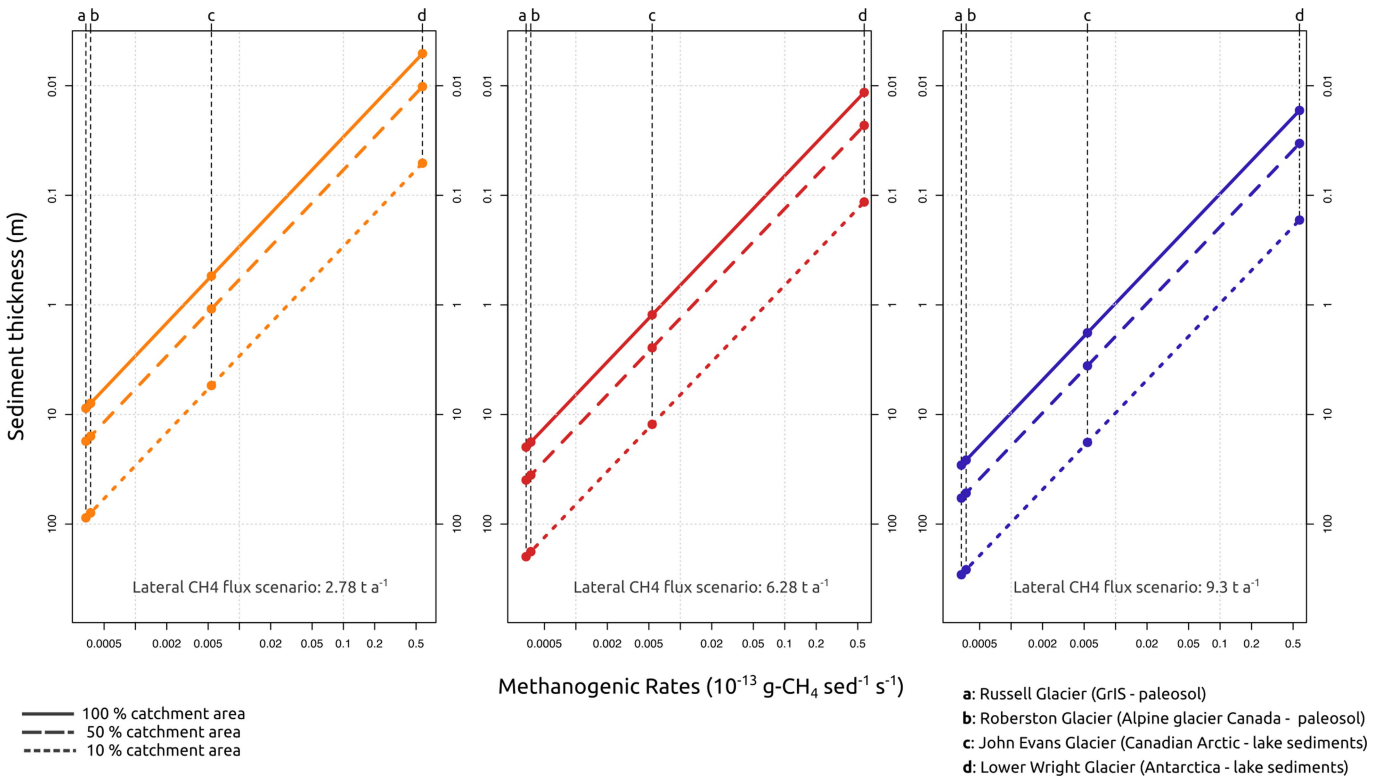
Extended Data Fig. 3 | Combination plot of $\delta^{13}C\text{-}CO_2$ and $\delta^{13}C\text{-}CH_4$ of LG runoff. Points denote $\delta^{13}C$ CO_2-CH_4 values for LG manual samples. Methanogenesis and microbial oxidation classification zones are derived and adapted from ref. [25].

© 2019 Springer Nature Limited. All rights reserved.



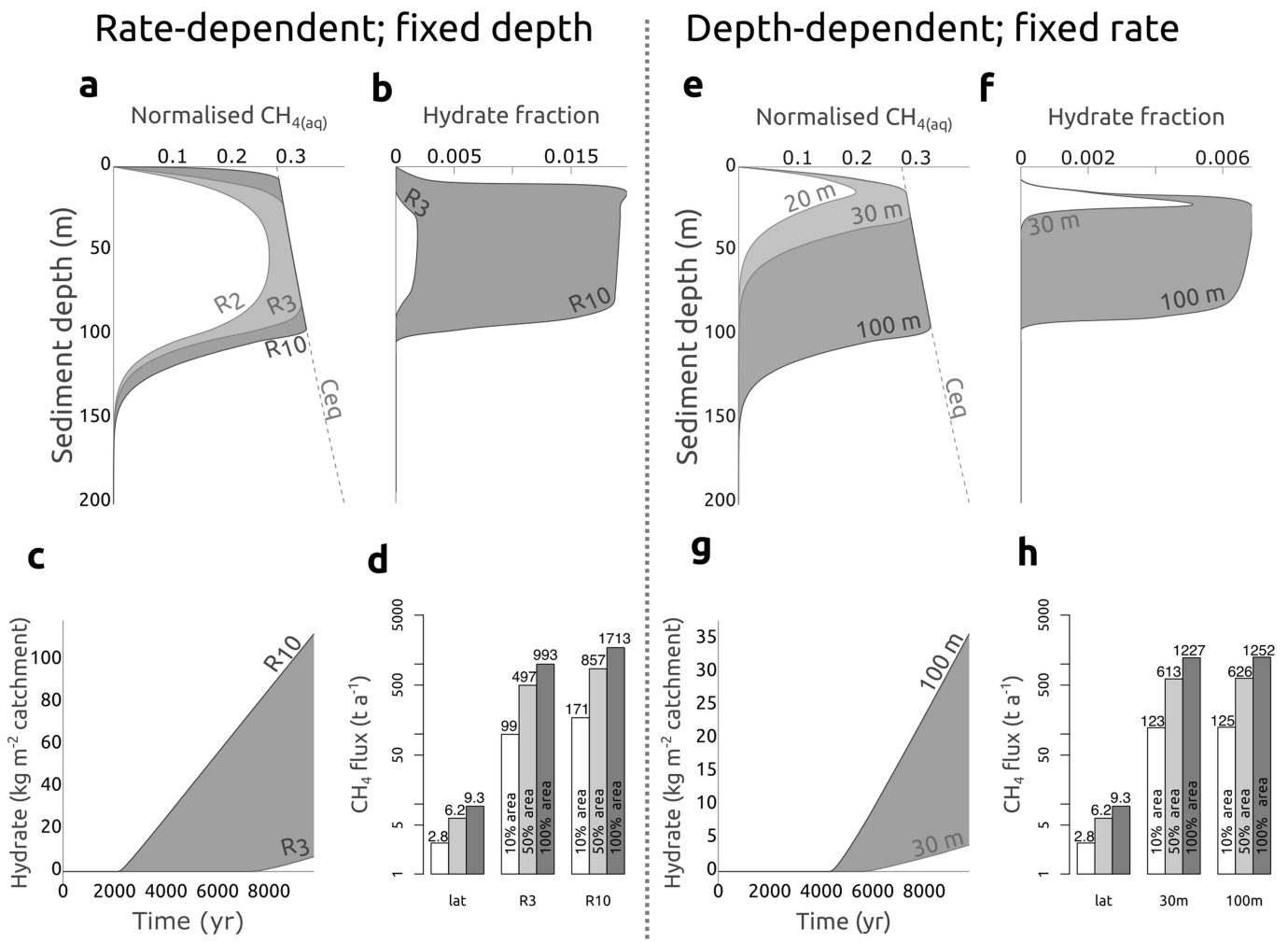
Extended Data Fig. 4 | LG 16S rRNA gene sequences related to methanotrophic and methanogenic clades. a, Relative abundance of the dominant operational taxonomical units related to bacterial methanotrophs (OTU00009) and archaeal methanogens. The box mid-lines represent medians; the IQR is represented by the lower and upper box boundaries and denote the 25th and 75th percentiles, respectively; whiskers indicate confidence intervals of 1.5 times the IQR, and points are outliers. **b**, **c**, Maximum-likelihood trees of 16S rRNA sequences related to methanotrophs rooted with the sequences of *Clostridium frigoriphilum* (**b**) and methanogens rooted with the sequences of *Acidibilus sulfurireducens* and *Caldisphaera draconis* (**c**).

© 2019 Springer Nature Limited. All rights reserved.



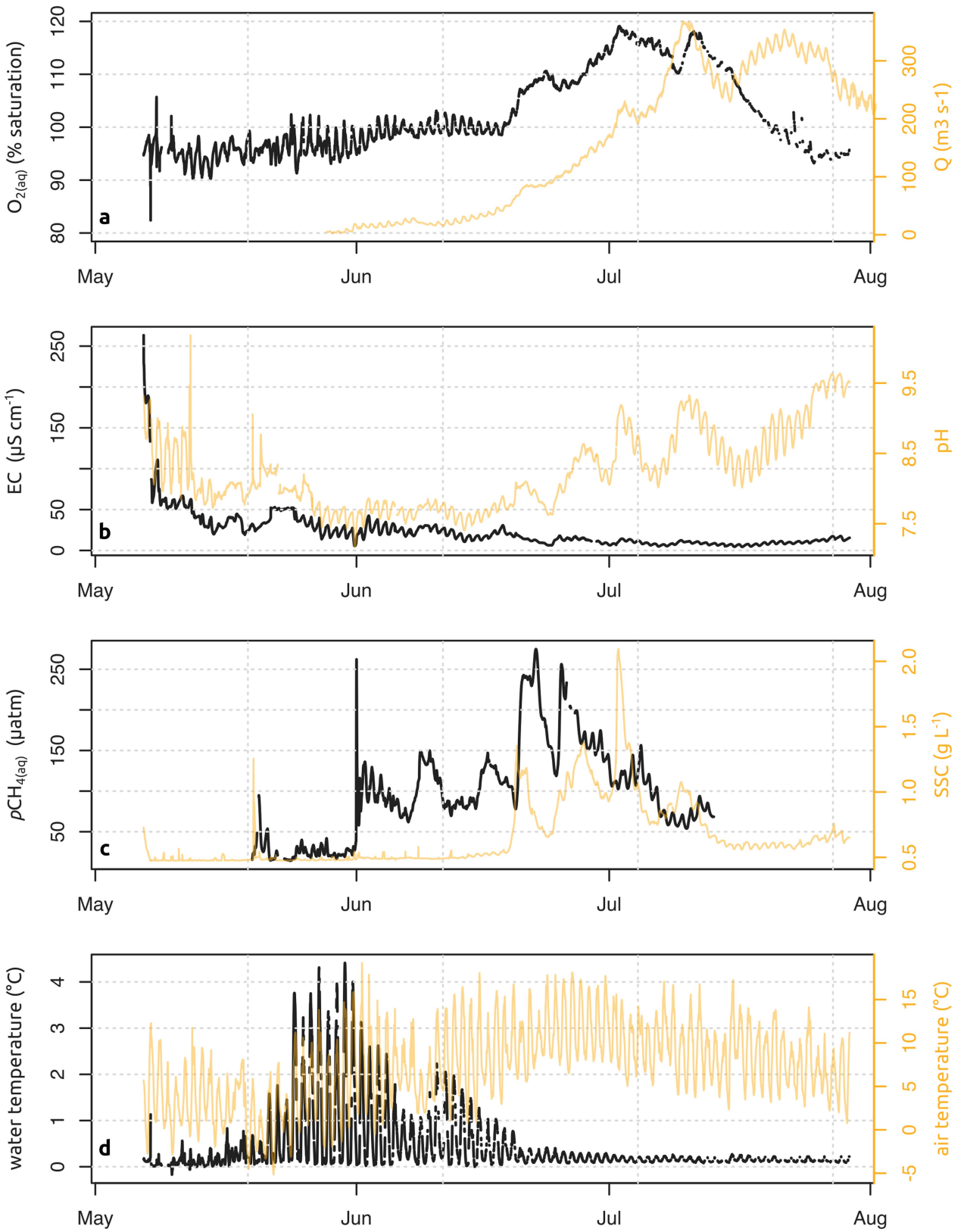
Extended Data Fig. 5 | Relationship between rates of subglacial methanogenesis, sediment thickness and observed annual CH_4 flux at LG. Each panel corresponds to the different yearly lateral $CH_{4(aq)}$ flux estimates measured in 2015 (see Fig. 2). Each line type corresponds to the sediment thickness required under different catchment area conditions: 100% (solid line), 50% (long-dashed line) or 10% (short-dashed line) of the subglacial catchment to contribute to the observed CH_4 flux. Any point on a line corresponds to the methanogenesis rate and subglacial sediment thickness required to generate the observed lateral CH_4 flux. The four points on each line correspond to known methanogenic rates recorded from different subglacial habitats[17].

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Summary of model conditions required for subglacial methane hydrate formation. a–d, Model results for a fixed methanogenic depth (100 m) but varying methanogenic rates (R_2 to R_{10}; that is, 2×10^{-15} to 10×10^{-15} grams CH_4 per gram of sediment per second). **e–h**, Outputs of model runs under a fixed methanogenic rate (5×10^{-15} grams CH_4 per gram of sediment per second) but varying methanogenic depths (20–100 m). **a**, **b**, **e**, **f**, Vertical profiles of methane solubility, dissolved methane and methane hydrates; methane concentrations are normalized to equilibrium concentration. **c**, **g**, Time required for methane hydrate formation under the modelled conditions. **d**, **h**, Diffusive CH_4 flux at the sediment–ice interface under methane hydrate conditions assuming three different catchment cover areas for methane hydrates (that is, 10%, 50% and 100% of the LG catchment), compared to the three lateral flux scenarios (**a**, **b**, **d**; Fig. 2); see Supplementary Information 2f.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Extended time series of geochemical measurements from the LG proglacial river. a–d, The EC, pH and SSC time series include those depicted in Fig. 1 but extend to measurements before and after the methane record. It should be noted that the $CH_{4(aq)}$ data in **c** are the HydroC partial pressure (in microatmospheres) measurements. Left and right vertical axes correspond to black and orange datasets, respectively.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | $CH_{4(aq)}$ concentration, fluxes and areal yield from LG, the GrIS and other world rivers

River	Catchment Area (10^6 km^2)	Discharge (km^3 a^{-1})	Concentration (μM $CH_{4(aq)}$)	Diffusive flux (mmol-CH_4 m^{-2} d^{-1})	Diffusive flux (Gg-CH_4 a^{-1})	Lateral flux (Gg-$CH_{4(aq)}$ a^{-1})	Areal Yield Diffusive (mmol-CH_4 m^{-2} a^{-1})	Areal Yield Lateral (mmol-$CH_{4(aq)}$ m^{-2} a^{-1})
LG runoff	**0.0006**	**1.45**	**0.27**	**14.94**	**-**	**0.0063**	**-**	**0.65**
GrIS runoffs†	-	*418* ‡	*0.27*	-	-	*2.1*	-	-
Lena Delta	2.49	821	0.07	9.22	150	0.9	3.8	0.02
Yukon Trib	0.19	48	0.75	3.16	12	0.6	4.0	0.19
Yukon MS	0.82	206	0.38	0.60	8	1.3	0.6	0.10
Negro Trib	0.69	455	0.87	-	-	6.0	-	1.29
Negro MS	0.69	1634	0.26	2.00	79	4.2	7.2	0.38
Solimões Trib	0.99	1985	0.33	-	-	10.6	-	1.17
Solimões MS	0.99	3507	0.06	1.60	160	3.3	10.1	0.21
Amazon	6.03	5444	0.18	0.92	490	15.5	5.1	0.16
Congo	3.71	1270	3.17	16.42	1906	64.6	32.1	1.10

Diffusive fluxes are calculated as grand means, except for the LG runoff diffusive flux, which corresponds to the medium flux scenario (scenario B in Fig. 2; see Methods for details). Except for the Amazon and Congo, lateral fluxes and yields are calculated using discharge-weighted means; see Supplementary Information 1c for references and calculation details.
†The GrIS-wide $CH_{4(aq)}$ flux was estimated using the LG discharge-weighted $CH_{4(aq)}$ concentration mean applied to the entire dataset of GrIS runoffs; this number is therefore speculative and was included for reference only.
‡From ref. [37]. Areal yields are for entire catchment areas, whereas diffusive fluxes refer to stream surface areas.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Stable-isotope details for CH_4 and CO_2

Sampling Time	$\delta^{13}C$-CH_4 (‰ VPDB)	$\delta^{2}H$-CH_4 (‰ VSMOW)	$\delta^{13}C$-CO_2 (‰ VPDB)
2015-05-04 12:00	-48.9	-	-16.9
2015-05-04 12:00	-46.1	-	-18.4
2015-05-30 11:30	-6.1	-	-12.2
2015-06-07 12:00	-61.3	-303	-15.3
2015-06-07 12:00	-53.1	-281	-16.0
2015-06-17 23:00	-56.7	-313	-15.8
2015-06-17 23:00	-55.2	-256	-14.5
2015-06-21 16:15	-59.3	-235	-22.2
2015-06-21 16:15	-57.6	-289	-24.1
2015-06-25 15:40	-63.4	-308	-14.6
2015-06-25 05:20	-56.6	-302	-14.8
2015-06-26 18:15	-60.2	-272	-23.6
2015-07-01 09:55	-55.9	-318	-26.8
2015-07-01 09:30	-55.2	-262	-26.2

The first three rows correspond to samples collected at the borehole and the chainsawed hole (see Supplementary Information 1b, Extended Data Fig. 3).

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 3 | Site-specific parameters applied in the one-dimensional hydrate model

Symbol	Parameter	Value	Unit
z_{xn}	Sediment thickness	20-100	m
h	Ice thickness	1,000	m
G	Geothermal gradient	0.025	°C m^{-1}
T(0)	Basal Temperature	-1	°C
sed	Sedimentation rate	0	m s^{-1}
vup	Upward fluid flux	0	m s^{-1}
φ0	Porosity	0.6	-
R_{xn}	Constant methane production rate	10^{-17} - 10^{-13}	g-CH_4 g^{-1} wet-sediment s^{-1}

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0771-1

Capture of nebular gases during Earth's accretion is preserved in deep-mantle neon

Curtis D. Williams[1]* & Sujoy Mukhopadhyay[1]

Evidence for the capture of nebular gases by planetary interiors would place important constraints on models of planet formation. These constraints include accretion timescales, thermal evolution, volatile compositions and planetary redox states[1–7]. Retention of nebular gases by planetary interiors also constrains the dynamics of outgassing and volatile loss associated with the assembly and ensuing evolution of terrestrial planets. But evidence for such gases in Earth's interior remains controversial[8–14]. The ratio of the two primordial neon isotopes, $^{20}Ne/^{22}Ne$, is significantly different for the three potential sources of Earth's volatiles: nebular gas[15], solar-wind-irradiated material[16] and CI chondrites[17]. Therefore, the $^{20}Ne/^{22}Ne$ ratio is a powerful tool for assessing the source of volatiles in Earth's interior. Here we present neon isotope measurements from deep mantle plumes that reveal $^{20}Ne/^{22}Ne$ ratios of up to 13.03 ± 0.04 (2 standard deviations). These ratios are demonstrably higher than those for solar-wind-irradiated material and CI chondrites, requiring the presence of nebular neon in the deep mantle. Furthermore, we determine a $^{20}Ne/^{22}Ne$ ratio for the primordial plume mantle of 13.23 ± 0.22 (2 standard deviations), which is indistinguishable from the nebular ratio, providing robust evidence for a reservoir of nebular gas preserved in the deep mantle today. The acquisition of nebular gases requires planetary embryos to grow to sufficiently large mass before the dissipation of the protoplanetary disk. Our observations also indicate distinct $^{20}Ne/^{22}Ne$ ratios between deep mantle plumes and mid-ocean-ridge basalts, which is best explained by addition of a chondritic component to the shallower mantle during the main phase of Earth's accretion and by subsequent recycling of seawater-derived neon in plate tectonic processes.

A long-standing debate about the formation of the Earth revolves around whether nebular gases were dissolved into a magma ocean during the early stages of accretion and preserved in the mantle to the present day[1–3,8–14]. Primordial neon isotopes (^{20}Ne and ^{22}Ne) can distinguish between different sources of volatiles in Earth's interior, as the three most likely sources (nebular gas, solar-wind-irradiated meteoritic materials and CI chondrites) have distinct $^{20}Ne/^{22}Ne$ ratios. If volatiles were acquired through the dissolution of nebular gases into a magma ocean on the proto-Earth[1–3,13,14], the mantle would be characterized by the nebular $^{20}Ne/^{22}Ne$ ratio of 13.36 ± 0.18 (2σ)[15]. On the other hand, if they were acquired primarily through the accretion of meteoritic material irradiated by solar wind[8–11], the mantle would be characterized by $^{20}Ne/^{22}Ne$ ratios of 12.52–12.75 (refs [9,16]). Dust grains in the nebula acquire a ratio of 12.52–12.75 through a combination of implantation of solar wind and erosion of the outer layers of the grain[8,9]. The proto-Earth might then have inherited this signature of solar wind implantation when the dust grains coalesced to form planetesimals, which subsequently merged to form planetary embryos. Finally, if volatiles in the proto-Earth were acquired mainly through accretion of CI chondrites, the mantle would be characterized by a $^{20}Ne/^{22}Ne$ ratio significantly lower than for nebular gas or solar-wind-irradiated materials, as the average CI chondrite value is 9.03 ± 2.46 (2σ)[17].

Accurately determining the source of neon in Earth's interior from the measurement of mantle-derived basalts is challenging because of pervasive syn-eruptive to post-eruptive atmospheric contamination. The $^{20}Ne/^{22}Ne$ ratio of Earth's atmosphere is 9.8 and probably reflects derivation of atmospheric neon from a chondritic source[12,18]. The $^{20}Ne/^{22}Ne$ ratio of Earth's mantle is higher than the atmospheric value, but atmospheric contamination lowers the measured $^{20}Ne/^{22}Ne$ ratios towards the atmospheric value. As a result, maximum measured $^{20}Ne/^{22}Ne$ ratios in basalts are frequently taken to represent the mantle value. Recent high-precision measurements of the $^{20}Ne/^{22}Ne$ ratios in mid-ocean-ridge basalts (MORBs) reach values of 12.48 ± 0.14 (2σ) (Extended Data Table 1), higher than for CI chondrites but similar to values in solar-wind-irradiated meteoritic material[8,9]. However, it is difficult to demonstrate that the maximum measured $^{20}Ne/^{22}Ne$ ratios in basalts represent the mantle value, completely free of syn-eruptive to post-eruptive air contamination.

Continental well-gases, such as those from New Mexico, USA, represent a different, yet complementary, repository of mantle fluids

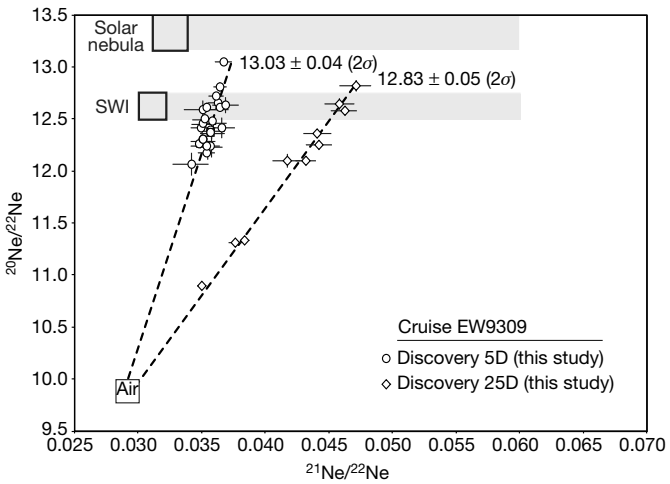


Fig. 1 | The neon isotopic composition for the individual step-crushes of the two plume-influenced MORBs. MORB samples were collected from the Discovery Ridge segment of the South Atlantic mid-ocean ridge. Error bars are 2σ. For reference, the neon isotopic compositions of air and solar nebula gas[15] and estimates of solar-wind-irradiated meteoritic materials[9,16] are shown. SWI represents the range of observed and modelled isotopic composition for solar-wind-irradiated materials. The grey shaded bars are extensions of the bounds on the $^{20}Ne/^{22}Ne$ ratios of the solar nebula and solar-wind-irradiated material, for comparison with the neon isotopic composition of oceanic basalts. The dashed lines through the step-crushes represent the mantle–air mixing lines. The slope of the line is determined by comparing the amount of nucleogenic ^{21}Ne to primordial ^{22}Ne, with steeper slopes indicating a higher proportion of primordial ^{22}Ne and therefore a less-degassed, more primitive mantle composition. At the 2σ level, our maximum measured $^{20}Ne/^{22}Ne$ ratios clearly exceed the values for solar-wind-irradiated materials.

[1]Department of Earth and Planetary Sciences, University of California-Davis, Davis, CA, USA. *e-mail: cdwill@ucdavis.edu

© 2019 Springer Nature Limited. All rights reserved.

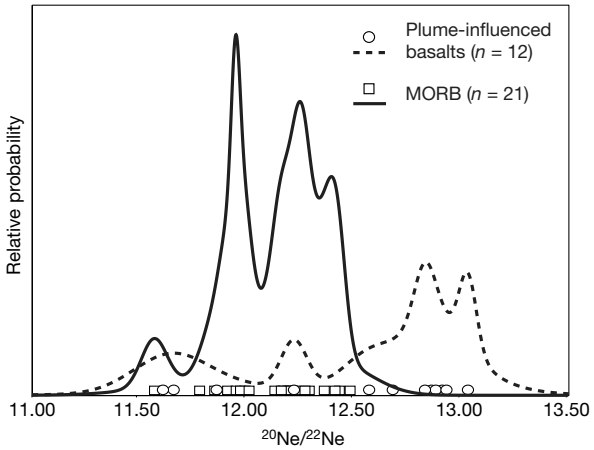


Fig. 2 | Relative probability functions for MORBs and plume-influenced basalts. Probability curves were constructed from the maximum measured neon isotopic composition ($^{20}Ne/^{22}Ne_{max}$) of globally distributed, mantle-derived samples from non-plume-influenced MORBs (solid curve) and plume-influenced basalts (dashed curve). Samples for which $^{20}Ne/^{22}Ne_{max}$ exceeds 11.5 with an associated 2σ uncertainty of at most 0.25 were selected to construct the probability curves. Also shown are the individual $^{20}Ne/^{22}Ne_{max}$ values as squares (MORBs) and circles (plume-influenced materials). To aid comparison between MORBs and plume-influenced basalts, only the highest $^{20}Ne/^{22}Ne$ value was selected from an individual mid-ocean-ridge segment and from a plume locality. There is a clear difference between the maximum measured $^{20}Ne/^{22}Ne$ ratios between non-plume-influenced MORBs and plume-influenced basalts, with non-plume-influenced MORBs displaying a sharp cut-off at a $^{20}Ne/^{22}Ne$ ratio of 12.5. The distinct peaks at lower $^{20}Ne/^{22}Ne$ ratio are not likely to be significant and are probably an artefact of the small dataset used to construct the relative probability curves (see also Extended Data Fig. 2). We note that after more than 30 years of neon isotopic measurements, not a single $^{20}Ne/^{22}Ne$ ratio in non-plume-influenced MORBs is observed to be resolvably higher than 12.49 ± 0.08 (2σ). On the other hand, several high-precision measurements of plume-influenced basalts display $^{20}Ne/^{22}Ne$ ratios that are resolvably higher than 12.6. Data sources are reported in Extended Data Table 1.

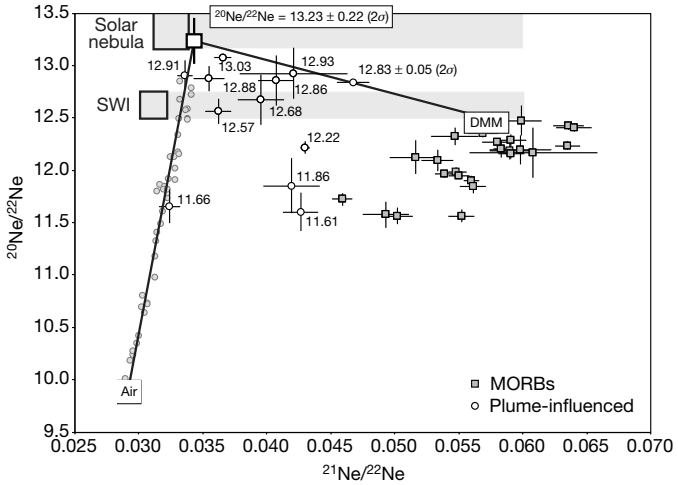


Fig. 3 | Determining the plume mantle $^{20}Ne/^{22}Ne$ ratio from two-component mixing arrays. Shown is the mixing array defined by the $^{20}Ne/^{22}Ne$ ratios of EW9309_25D (characterized by a $^{20}Ne/^{22}Ne$ ratio of 12.83 ± 0.05 (2σ)) and the depleted MORB mantle (DMM)[19] that is projected back to the Galápagos plume–air mixing line[20]. SWI, solar-wind-irradiated material; grey shaded bars as in Fig. 1. We chose Galápagos for this exercise as it represents the most primitive mantle plume in terms of the nucleogenic neon ($^{21}Ne/^{22}Ne$) isotopic composition[20]. Also shown are the highest measured $^{20}Ne/^{22}Ne$ ratios for globally distributed, mantle-derived MORBs (squares) and plume-influenced basalts (circles) used to construct Fig. 2. Note that the high-precision MORB data only reach the lower end of the solar-wind-irradiated range whereas several high-precision plume-influenced basalt data exceed the SWI range. Because the $^{20}Ne/^{22}Ne$ ratio in non-plume-influenced MORBs and plume-influenced basalts is distinct (Fig. 2), samples that reflect a mixture of plume and depleted MORB mantles, such as the Discovery samples[22,24], must be characterized by $^{20}Ne/^{22}Ne$ ratios that are intermediate between the MORB mantle value and the most primitive plume mantle value. The mixing line in this space is linear, and the intersection point of the mixing line with the Galápagos–air line occurs at a $^{20}Ne/^{22}Ne$ ratio of 13.23 ± 0.22 (2σ). This value represents the composition of the most primitive plume mantle in the present-day Earth and is indistinguishable from the nebular $^{20}Ne/^{22}Ne$ ratio.

from those trapped in basalts. These well-gases represent a mixture of crustal fluids contaminated with air to which mantle fluids are added[10,19]. The mixing systematics allows the present-day $^{20}Ne/^{22}Ne$ ratio of MORB mantle to be constrained to 12.49 ± 0.08 (2σ)[19], similar to the highest measured values in MORBs.

On the other hand, several plume-influenced locations display measured $^{20}Ne/^{22}Ne$ ratios that exceed the (non-plume-influenced) MORB mantle value (see Methods for definition of plume-influenced locations). These high $^{20}Ne/^{22}Ne$ ratios include measurements from the Kola Peninsula of Russia ($^{20}Ne/^{22}Ne = 13.04 \pm 0.40$ (2σ))[13], Iceland (12.88 ± 0.12 (2σ))[14], the Galápagos islands (12.91 ± 0.14 (2σ))[20], the 17° S anomaly (12.86 ± 0.24 (2σ))[21] and plume-influenced MORBs in the south Atlantic (13.10 ± 0.50 (2σ))[22]. The relatively high ratios may indicate the presence of a nebular component in the deep mantle[13,14]. However, given the associated analytical uncertainties, these plume ratios may also be consistent with solar-wind-irradiated material[9], with the lower values in MORBs resulting from the subduction of atmospheric neon[9,23]. Furthermore, recent measurements of individual vesicles in basalts from the Galápagos plume were used to interpret the high $^{20}Ne/^{22}Ne$ ratios in plumes as arising from mass-dependent fractionation and therefore not being representative of the neon composition of the plume mantle[8]. Instead, a $^{20}Ne/^{22}Ne$ ratio of 12.65 ± 0.08 (2σ) was advocated for the plume mantle[8]. This value for the plume mantle would suggest a near-uniform $^{20}Ne/^{22}Ne$ ratio for the whole mantle that is similar to values in meteoritic material irradiated by solar wind.

The ongoing debate associated with the measurement and interpretation of $^{20}Ne/^{22}Ne$ ratios in plume-derived materials[8,9,13,14,23] precludes a reliable assessment of (1) whether the neon in the plume mantle is characterized by nebular gas or solar-wind-irradiated material, (2) whether there are significant differences between present-day MORB and plume mantles, and (3) whether the mantle ratio has changed over time because of late accretion and/or subduction of atmospheric neon. Resolving these debates is essential, as acquisition of volatiles through dissolution of nebular gases into a magma ocean represents a fundamentally different mechanism of volatile accretion from accretion of irradiated dust grains. These two modes of volatile accretion also provide distinctly different constraints on physical processes operating in the early Solar System[2,9,13]. Towards that end, we present new high-precision neon isotope measurements from plume-influenced samples in the south Atlantic (Supplementary Table 1).

Figure 1 displays the measured neon isotopic composition for individual step-crushes of two samples influenced by the Discovery plume in the south Atlantic, with $^{20}Ne/^{22}Ne$ ratios reaching 12.83 ± 0.05 (2σ) and 13.03 ± 0.04 (2σ). In both samples, individual step-crushes display strong linear correlations that reflect two-component mixing between atmospheric neon and mantle neon. The strong linearity of $^{20}Ne/^{22}Ne$ versus $^{21}Ne/^{22}Ne$ indicates that mass-dependent fractionation processes do not have a role in generating the observed high $^{20}Ne/^{22}Ne$ ratios. Furthermore, $^{4}He/^{3}He$ and $^{38}Ar/^{36}Ar$ ratios measured for the same crushing steps rule out mass-dependent fractionation generating the observed $^{20}Ne/^{22}Ne$ ratios during bubble formation (Extended Data Fig. 1). Therefore, the measured neon isotope ratios (Fig. 1) represent two-component mixing between unfractionated mantle neon and a syn-eruptive to post-eruptive atmospheric contaminant. Because the highest measured value is not necessarily entirely free of syn-eruptive

© 2019 Springer Nature Limited. All rights reserved.

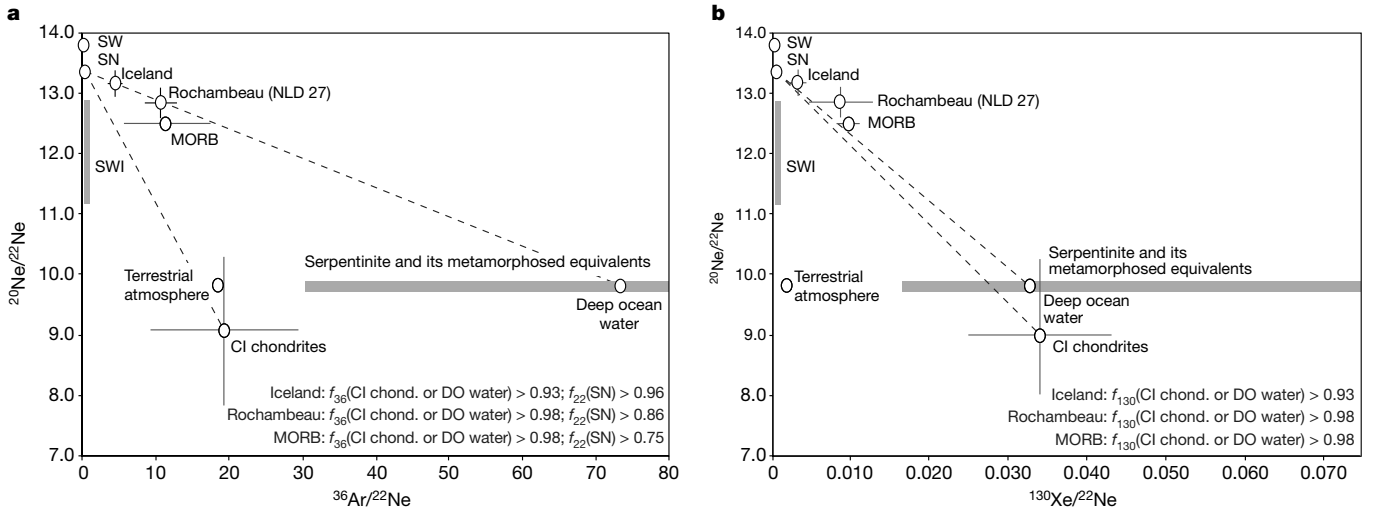


Fig. 4 | The $^{36}Ar/^{22}Ne$–$^{20}Ne/^{22}Ne$ and $^{130}Xe/^{22}Ne$–$^{20}Ne/^{22}Ne$ systematics of the MORB mantle and plume-influenced basalts from Iceland and Rochambeau (which samples the Samoan plume). Also shown are solar-wind data collected during the Genesis mission (SW), estimates of the solar nebula (SN) and CI chondrites. The data points and end-member compositions are reported in Extended Data Table 2. Error bars on the $^{20}Ne/^{22}Ne$ ratios are 2σ, while the errors bars on the $^{36}Ar/^{22}Ne$ and $^{130}Xe/^{22}Ne$ ratios are only 1σ for clarity. The grey shaded region represents the range of observed and modelled isotopic compositions for solar wind implantation into dust (abbreviated SWI). Both the MORB and mantle plume compositions fall on the linear trend defined by mixing of nebular gases with deep ocean water and/or chondritic gases. The fractions of recycling ^{22}Ne, ^{36}Ar and ^{130}Xe are represented by f_{22}, f_{36} and f_{130}, respectively. These relationships strongly suggest that late accretion of chondritic volatiles and recycling of seawater-derived neon into the MORB mantle has lowered its $^{20}Ne/^{22}Ne$ ratio. In addition, the linear trend displayed by the mantle compositions would be highly fortuitous if the mantle composition were similar to solar-wind-irradiated material, as the mixing lines between solar-wind-irradiated materials and chondritic/ocean water signatures do not pass through the mantle compositions.

to post-eruptive atmospheric contamination, the measured $^{20}Ne/^{22}Ne$ ratios of 12.83 ± 0.05 (2σ) and 13.03 ± 0.04 (2σ) are robust minimum mantle values for these plume-influenced basalts.

These minimum $^{20}Ne/^{22}Ne$ ratios measured in plume-influenced basalts (Fig. 1) are resolvably higher than the MORB mantle $^{20}Ne/^{22}Ne$ ratio of 12.49 ± 0.08 (2σ)[19] as well as solar-wind-irradiated meteoritic materials, which should have $^{20}Ne/^{22}Ne$ ratios of 12.52–12.75 (refs [8–11,16]). Implantation of solar wind into dust grains may yield higher $^{20}Ne/^{22}Ne$ ratios (up to 12.89) if implantation and sputtering do not reach steady state[9]. However, chondrules that were irradiated before their incorporation into their meteorite parent bodies have $^{20}Ne/^{22}Ne$ ratios less than 12.5 and are more often characterized by a galactic cosmic-ray signature ($^{20}Ne/^{22}Ne \approx 1$) (see Methods). Because chondrules were formed by melting of nebular dust, these observations demonstrate either that the dust did not acquire $^{20}Ne/^{22}Ne$ ratios greater than 12.5 through solar wind implantation or that signatures of solar wind implantation are not preserved through chondrule melting processes or subsequent formation and evolution of planetesimals (see Methods). Given this, solar wind implantation into nebular dust cannot be responsible for the high $^{20}Ne/^{22}Ne$ ratios observed in plumes. Rather, our measured $^{20}Ne/^{22}Ne$ ratios of up to 13.03 ± 0.04 (2σ) require the presence of nebular neon in the present-day plume mantle.

Along with recent high-precision measurements, our new analyses indicate that the MORB and the plume mantles are characterized by two discrete $^{20}Ne/^{22}Ne$ ratios. Figure 2 illustrates that the distributions of the maximum measured $^{20}Ne/^{22}Ne$ ratios in MORBs and plumes are statistically distinct, with both the mode and the tail of the plume distribution shifted to higher $^{20}Ne/^{22}Ne$ ratios (see also Extended Data Fig. 2). For MORBs, the values tail off at a $^{20}Ne/^{22}Ne$ ratio of about 12.5, consistent with the value of 12.49 ± 0.08 (2σ) determined for the depleted MORB mantle[19]. On the other hand, plume-influenced basalts display a major mode at 12.9 with one tail of the distribution trending towards higher values (Fig. 2). The different $^{20}Ne/^{22}Ne$ ratios of plumes and MORBs cannot result from processes related to eruption depths, as the eruption depths completely overlap for the two data sets (Extended Data Fig. 3). Nor are the distinct $^{20}Ne/^{22}Ne$ ratios for MORBs and plumes due to nucleogenic ingrowth (see Methods). Rather, the two discrete maximum modes require that Earth's present-day mantle contains at least two distinct accretionary sources of neon.

The observation of nebular neon in the plume mantle is reinforced by considering that the neon isotope ratios measured in our samples represent a two-component mixture of a plume mantle with the MORB mantle[22,24]. For example, the basalts influenced by the Discovery plume display two nearly orthogonal trends in Sr–Pb isotope space that represent mixing between the depleted MORB mantle and a plume mantle[24]. Furthermore, globally, in $^{21}Ne/^{22}Ne$ versus $^{4}He/^{3}He$ space, all plume and plume-influenced basalts fall on a single hyperbolic mixing line between a depleted mantle and the least degassed plume mantle[25]. These observations of two-component mixing indicate that the measured $^{20}Ne/^{22}Ne$ ratios of 12.83 ± 0.05 (2σ) and 13.03 ± 0.04 (2σ) in the Discovery Ridge samples must be intermediate between the MORB mantle value of 12.49 ± 0.08 (2σ)[19] and the $^{20}Ne/^{22}Ne$ ratio of the least degassed plume mantle. The $^{20}Ne/^{22}Ne$ ratio of the least degassed plume mantle can be determined by projecting a two-component mixing line from the depleted MORB mantle[19] to the Galápagos–air mixing line[20] such that the maximum measured values in plume-influenced basalts lie on or below this two-component mixing line (Fig. 3). Extrapolation to the Galápagos–air mixing line is made because samples from the Galápagos represent the most primitive $^{21}Ne/^{22}Ne$ ratios measured in plume-influenced basalts so far. Figure 3 shows that the intersection of the plume–MORB mixing line with the Galápagos–air mixing line defines the least degassed plume mantle $^{20}Ne/^{22}Ne$ ratio to be 13.23 ± 0.22 (2σ). This plume mantle ratio is indistinguishable from the nebular $^{20}Ne/^{22}Ne$ ratio of 13.36 ± 0.18 (2σ)[15], providing further evidence for the preservation of nebular neon in the present-day plume mantle.

The presence of nebular neon in the plume mantle requires acquisition of other nebular volatiles such as hydrogen, carbon and nitrogen by the proto-Earth. This statement may seem paradoxical given evidence for chondritic xenon[26] as well as chondritic hydrogen, carbon and nitrogen isotopic compositions in plume-influenced samples[12,18]. However, late accretion of chondritic volatiles and preferential recycling of seawater-derived volatiles into the mantle may overprint the nebular xenon, hydrogen, carbon and nitrogen signature in these two-component mixtures. This is demonstrated in Fig. 4, where the primordial

© 2019 Springer Nature Limited. All rights reserved.

$^{36}Ar/^{22}Ne$ and $^{130}Xe/^{22}Ne$ ratios are plotted against $^{20}Ne/^{22}Ne$ ratio for both plume-influenced basalts and MORBs. Plume-influenced and MORB samples trend from a nebular-like composition towards a component similar to CI chondrite[17] and/or seawater[11,19]. The correlations observed in Fig. 4 show that for measured $^{20}Ne/^{22}Ne$ ratios in plume-influenced basalts, more than 96% of the ^{22}Ne would be derived from nebular gas, while more than 93% of the ^{36}Ar and ^{130}Xe would be derived from a chondritic and/or seawater source. The lack of a nebular signature in hydrogen, carbon and nitrogen isotopic compositions may also reflect the sequestration of these elements into the core followed by later addition of chondritic material to the growing Earth. Unlike neon, experimental studies indicate that carbon, nitrogen and hydrogen are siderophile under core formation conditions[6,7] (although metal–silicate partitioning for hydrogen is still debated[27]) and could, therefore, be strongly partitioned into the Fe-metal relative to the silicate magma ocean.

The $^{20,22}Ne$–^{36}Ar–^{130}Xe systematics (Fig. 4) suggest that neon in the MORB mantle can be derived from the plume mantle by the addition of a chondritic component during the later (post-nebula dissipation) stages of Earth's accretion. The chondritic neon may also have been added to the mantle through plate tectonic recycling of seawater-derived noble gases in hydrous minerals[23]. Because the mantle noble gases lie between nebular and chondritic/ocean-water compositions (Fig. 4), MORB mantle neon cannot be derived exclusively from solar-wind-irradiated materials; for example, a mixing line between solar-wind-irradiated material with $^{20}Ne/^{22}Ne$ ratios less than 12.5 and chondritic/ocean water would not pass through the mantle compositions.

The capture and preservation of nebular gases provides constraints on growth rates of terrestrial planets. Astronomical observations indicate that nebular gas typically disperses within an e-folding timescale of 2.5 million years[28]. Thus, the presence of nebular neon requires the proto-Earth to have reached a sufficient mass within a few million years to capture nebular volatiles and dissolve them into a magma ocean. This scenario is consistent with embryos in the terrestrial-planet-forming region growing to Mars-size, and potentially larger, within 2 million years[29,30]. In addition, planet formation at about 1 astronomical unit in a gas-rich, nebular environment has been directly observed using the Atacama Large Millimeter Array[31]. Our new neon isotopic data from the deep mantle, in combination with previous measurements[13,14,20], suggest a similar environment for the proto-Earth[1–3,13,14]. Therefore, the capture of nebular gases could be a common feature associated with the embryo stage of terrestrial planet formation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0771-1.

Received: 28 May 2018; Accepted: 24 September 2018;
Published online 5 December 2018.

1. Mizuno, H., Nakazawa, K. & Hayashi, C. Dissolution of the primordial rare gases into the molten Earth's material. *Earth Planet. Sci. Lett.* **50**, 202–210 (1980).
2. Sasaki, S. & Nakazawa, K. Did a primary solar-type atmosphere exist around the proto-Earth? *Icarus* **85**, 21–42 (1990).
3. Harper, C. L. Jr & Jacobsen, S. B. Noble gases and Earth's accretion. *Science* **273**, 1814–1818 (1996).
4. Matsui, T. & Abe, Y. Evolution of an impact-induced atmosphere and magma ocean on the accreting Earth. *Nature* **319**, 303–305 (1986).
5. Kuramoto, K. & Matsui, T. Partitioning of H and C between the mantle and core during the core formation in the Earth: its implications for the atmospheric evolution and redox state of early mantle. *J. Geophys. Res.* **101**, 14909 (1996).
6. Hirschmann, M. M., Withers, A. C., Ardia, P. & Foley, N. T. Solubility of molecular hydrogen in silicate melts and consequences for volatile evolution of terrestrial planets. *Earth Planet. Sci. Lett.* **345–348**, 38–48 (2012).
7. Hirschmann, M. M. Magma ocean influence on early atmosphere mass and composition. *Earth Planet. Sci. Lett.* **341**, 48–57 (2012).
8. Péron, S., Moreira, M., Putlitz, B. & Kurz, M. D. Solar wind implantation supplied light volatiles during the first stage of Earth accretion. *Geochem. Perspect. Lett.* **3**, 151–159 (2017).
9. Moreira, M. & Charnoz, S. The origin of the neon isotopes in chondrites and on Earth. *Earth Planet. Sci. Lett.* **433**, 249–256 (2016).
10. Ballentine, C. J., Marty, B., Lollar, B. S. & Cassidy, M. Neon isotopes constrain convection and volatile origin in the Earth's mantle. *Nature* **433**, 33–38 (2005).
11. Trieloff, M., Kunz, J., Clague, D. A., Harrison, D. & Allègre, C. J. The nature of pristine noble gases in mantle plumes. *Science* **288**, 1036–1038 (2000).
12. Marty, B. The origins and concentrations of water, carbon, nitrogen and noble gases on Earth. *Earth Planet. Sci. Lett.* **313–314**, 56–66 (2012).
13. Yokochi, R. & Marty, B. A determination of the neon isotopic composition of the deep mantle. *Earth Planet. Sci. Lett.* **225**, 77–88 (2004).
14. Mukhopadhyay, S. Early differentiation and volatile accretion recorded in deep-mantle neon and xenon. *Nature* **486**, 101–104 (2012).
15. Heber, V. S. et al. Isotopic mass fractionation of solar wind: evidence from fast and slow solar wind collected by the Genesis mission. *Astrophys. J.* **759**, 121 (2012).
16. Black, D. C. On the origins of trapped helium, neon and argon isotopic variations in meteorites—II. Carbonaceous meteorites. *Geochim. Cosmochim. Acta* **36**, 377–394 (1972).
17. Mazor, E., Heymann, D. & Anders, E. Noble gases in carbonaceous chondrites. *Geochim. Cosmochim. Acta* **34**, 781–824 (1970).
18. Halliday, A. N. The origins of volatiles in the terrestrial planets. *Geochim. Cosmochim. Acta* **105**, 146–171 (2013).
19. Holland, G. & Ballentine, C. J. Seawater subduction controls the heavy noble gas composition of the mantle. *Nature* **441**, 186–191 (2006).
20. Péron, S. et al. Neon isotopic composition of the mantle constrained by single vesicle analyses. *Earth Planet. Sci. Lett.* **449**, 145–154 (2016).
21. Kurz, M. D. et al. Correlated helium, neon, and melt production on the super-fast spreading East Pacific Rise near 17° S. *Earth Planet. Sci. Lett.* **232**, 125–142 (2005).
22. Sarda, P., Moreira, M., Staudacher, T., Schilling, J.-G. & Allègre, C. J. Rare gas systematics on the southernmost Mid-Atlantic Ridge: constraints on the lower mantle and the Dupal source. *J. Geophys. Res.* **105**, 5973–5996 (2000).
23. Kendrick, M. A., Scambelluri, M., Honda, M. & Phillips, D. High abundances of noble gas and chlorine delivered to the mantle by serpentinite subduction. *Nat. Geosci.* **4**, 807 (2011).
24. Douglass, J., Schilling, J. G. & Fontignie, D. Plume-ridge interactions of the Discovery and Shona mantle plumes with the southern Mid-Atlantic Ridge (40°–55° S). *J. Geophys. Res. Solid Earth* **104**, 2941–2962 (1999).
25. Moreira, M. Noble gas constraints on the origin and evolution of Earth's volatiles. *Geochem. Perspect.* **2**, 229–403 (2013).
26. Caracausi, A., Avice, G., Burnard, P. G., Füri, E. & Marty, B. Chondritic xenon in the Earth's mantle. *Nature* **533**, 82–85 (2016).
27. Clesi, V. et al. Low hydrogen contents in the cores of terrestrial planets. *Sci. Adv.* **4**, e1701876 (2018).
28. Mamajek, E. E., Usuda, T., Tamura, M. & Ishii, M. Initial conditions of planet formation: lifetimes of primordial disks. *AIP Conf. Proc.* **1158**, 3–10 (2009).
29. Dauphas, N. & Pourmand, A. Hf–W–Th evidence for rapid growth of Mars and its status as a planetary embryo. *Nature* **473**, 489–492 (2011).
30. Rubie, D. C. et al. Accretion and differentiation of the terrestrial planets with implications for the compositions of early-formed Solar System bodies and accretion of water. *Icarus* **248**, 89–108 (2015).
31. Andrews, S. M. et al. Ringed substructure and a gap at 1 au in the nearest protoplanetary disk. *Astrophys. J. Lett.* **820**, L40 (2016).

Acknowledgements This work is supported by an NSF EAR Postdoctoral Fellowship and NSF grant EAR-1250419. Discovery samples were obtained during the cruise EW9309 of RV *Maurice Ewing* (in November–December 1993) and provided by the Marine Geological Samples Laboratory, of the Graduate School of Oceanography, University of Rhode Island.

Reviewer information *Nature* thanks D. Graham and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Both authors contributed to the design of the study, analyses and data processing. C.D.W. wrote the manuscript with input from S.M.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0771-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0771-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.D.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

Analytical methods. Basaltic glass chipped from pillow lavas was leached in dilute nitric acid, rinsed in acetone and then dried. Two to five grams of glass were loaded into a stainless-steel piston crusher, baked at 90–100 °C for 48 h and then pumped for an additional 7 days until blanks were low and stable. To release magmatic gases trapped in vesicles, samples were step-crushed under ultra-high vacuum using a hydraulic ram in the UC Davis Noble Gas Laboratory. Sequential exposure to hot and cold getters (SAES) removed active gases. The noble gases were then trapped (except for He) on a cryogenic cold finger. Helium was separated from neon at 32 K and let into the mass spectrometer (Nu Noblesse). The measurements were carried out at 200 μA trap current and an electron accelerating voltage of 60 eV. We measured ^{4}He with a Faraday cup and ^{3}He with an on-axis discrete dynode multiplier operating in pulse counting mode and fitted with an energy filter to reduce scattered ions. Following the helium measurements, neon was released from the cryogenic cold finger at 74 K. Neon was measured in multi-collection mode using three discrete dynode multipliers. We measured ^{21}Ne with the axial multiplier, and ^{20}Ne and ^{22}Ne with the low- and high-mass multipliers, respectively. An automated liquid nitrogen trap was used to keep the argon and carbon dioxide backgrounds low. Corrections for isobaric interference from doubly charged argon ($Ar^{2+}/Ar^{+} = 0.082 \pm 0.002$) and carbon dioxide ($CO_2^{2+}/CO_2^{+} = 0.0054 \pm 0.0006$) were made, and were typically less than 2% on the ^{20}Ne/^{22}Ne ratio. Following the neon measurements, argon was released from the cryogenic cold finger at 185 K. The argon isotopes were measured in multi-collection mode with ^{40}Ar measured with a Faraday detector, and ^{38}Ar and ^{36}Ar were measured on the axial and low-mass multipliers, respectively.

Procedural blanks measurements were typically less than 0.01% of the measured sample ^{4}He signal, <2% of the measured sample ^{22}Ne signal and <10% of the measured sample ^{36}Ar signal. Each step-crush was bracketed by blank measurements to ensure that blanks were stable. With the exception of helium, blanks are atmospheric in composition. Mass discrimination and sensitivity for neon and argon were determined with air standards and with the HH3 standard for helium[32]. Instrumental drift was monitored through sample-standard bracketing with additional standards run overnight (Extended Data Fig. 4). Uncertainties in gas abundances and isotope ratios for each step-crush are determined by propagating the uncertainties associated with both individual sample measurements and the reproducibility of analysed air standards of comparable signal size, as well as blank measurements that bracket each individual sample measurements using the general formula for error propagation[33]. Because blank corrections for Ne are <2% of the sample signals, they have little to no effect on the reported ratios and do not affect any of our conclusions; the analytical errors are almost entirely dominated by the external reproducibility of the air standards (see Extended Data Fig. 4). Noble gas abundances and isotope ratios are reported in Supplementary Table 1.

Literature compilation and generation of Fig. 2 and Extended Data Fig. 2. Plume-influenced locations were defined based on the plume catalogue developed by ref. [34], are within 500 km of a plume[35] or have a helium isotopic signature more than 3σ away from the mean value for MORBs. Literature data were compiled from locations worldwide where at least one ^{20}Ne/^{22}Ne ratio exceeds 11.5 with an associated uncertainty equal to or less than 0.25 at the 2σ level[36–50]. This filter results in a compilation of data that allows us to discriminate whether Earth's mantle comprises one, or more, discrete reservoirs of neon: for example, between non-plume-influenced MORBs and plume-influenced basalts. From this new compilation, only the highest ^{20}Ne/^{22}Ne value was selected from an individual mid-ocean-ridge segment or from an individual plume location to construct Fig. 2. The selection of only the highest ^{20}Ne/^{22}Ne value was done to alleviate oversampling a particular location when constructing the frequency diagrams. All of the literature data used in this study are reported in Extended Data Table 1.

Figure 2 represents a relative probability plot, which is constructed by summing several Gaussian distributions whose mean values and standard deviations correspond to the mean value of each individual measurement and their respective analytical uncertainties (2σ). This is done to reduce the importance of less precise measurements and emphasize more precise measurements. Note that the distinct peaks at lower ^{20}Ne/^{22}Ne ratios are not likely to be significant as they are probably an artefact of the small data set used to construct the relative probability curves. Kernel density estimates were also calculated using the Matlab Curve Fitting Toolbox with bandwidths of 0.16 and 0.12 for MORBs and plume-influenced materials, respectively (see Extended Data Fig. 2). The kernel density estimator results in slightly broader distributions than observed in Fig. 2 but does not change the main conclusions that there is a clear difference between the maximum measured ^{20}Ne/^{22}Ne ratios between non-plume-influenced MORBs and plume-influenced basalts, with MORBs displaying a sharp cut-off at a ^{20}Ne/^{22}Ne ratio of 12.5.

Determining the most primitive plume mantle ^{20}Ne/^{22}Ne ratio. For calculating the slope of the Galápagos–air mixing line, the Galápagos data points were translated such that the atmospheric neon composition was at (0, 0). The y data were then scaled by the square root of the ratio of the variance in x to the variance in y to put x and y on the same scale and then fitted with an equation of the form $y = mx$. The x and y error-weighted best-fit slope was computed by minimizing the value of χ^2. The uncertainty in the slope was calculated with a Monte Carlo method; the x and y data were varied at random to re-compute the fit, from which the confidence limit on the best fit was calculated. The plume–MORB mixing line is a two-point straight line computed by passing the line through the MORB mantle estimate (12.49 ± 0.08 (2σ)) and the Discovery sample data point 12.83 ± 0.05 (2σ). The choice of the Discovery data point was dictated by the requirement that the highest measured values in plume-influenced basalts that are intermediate between the most primitive plume mantle and the MORB mantle in ^{20}Ne/^{22}Ne–^{21}Ne/^{22}Ne space lie either on or below the mixing line. The uncertainty on the slope and intercept of the straight line was computed by randomly varying the MORB and Discovery data points. The uncertainties in both the Galápagos line and the plume–MORB mixing line were propagated in computing the uncertainty of the neon isotopic composition at the intersection point of the two lines.

Sensitivity of ^{20}Ne/^{22}Ne ratios to nucleogenic ingrowth. While the well-recognized production of nucleogenic ^{21}Ne decreases the mantle ^{21}Ne/^{22}Ne over time, the production of nucleogenic ^{20}Ne and ^{22}Ne may have also lowered the mantle ^{20}Ne/^{22}Ne over time. The sensitivity of ^{20}Ne/^{22}Ne ratios to nucleogenic ingrowth can be investigated by using the reported nucleogenic production ratios[51–55] and assuming that the reservoir has evolved as a closed system over 4.5 billion years of Earth history. The mantle production ratios for ^{21}Ne/^{22}Ne and ^{20}Ne/^{22}Ne range from 32.0 to 128.2 and from 2.5 to 12.4, respectively[51–55], with the lowest production ratios[55] producing the largest shift in ^{20}Ne/^{22}Ne ratios. Using the lowest reported production ratios[55], the ^{20}Ne/^{22}Ne ratio of the mantle decreases by only 0.01 as the mantle ^{21}Ne/^{22}Ne evolves from the nebular ^{21}Ne/^{22}Ne ratio of 0.0328 to the MORB mantle ^{21}Ne/^{22}Ne ratio[19] of 0.0578 ± 0.0006 (2σ). Therefore, evolution of a nebular reservoir to the present-day plume ^{21}Ne/^{22}Ne ratios, such as observed at Galapagos or for the Discovery samples reported here, would have decreased the ^{20}Ne/^{22}Ne ratios in the third and fourth decimal places (^{20}Ne/^{22}Ne = 13.3597–13.3552). These shifts are negligible as they are significantly lower than the reported analytical uncertainties.

Crustal fluids and natural gases, however, indicate that the production of ^{22}Ne may be underestimated in the crust[51,56,57]. The crustal ^{21}Ne/^{22}Ne production ratio may be 0.52, a factor of about 7 times lower than the theoretical production ratio of 4; this lower ^{21}Ne/^{22}Ne production ratio may be a result of the close association of uranium and thorium with fluorine in crustal lithologies[56,57]. If this close association of uranium and thorium with fluorine holds for the mantle, then the production of ^{22}Ne in the mantle may also be underestimated. The impact of a potentially lower mantle production ratios is investigated here by scaling the reported mantle production ratios[51–55] by the same correction factor as for crustal materials (that is, a factor of 7). Scaling the most extreme ^{21}Ne/^{22}Ne and ^{20}Ne/^{22}Ne mantle production ratios[53] of 32 and 2.5, respectively, results in corrected mantle productions ratios of 4.19 and 0.33 for ^{21}Ne/^{22}Ne and ^{20}Ne/^{22}Ne, respectively. Using these corrected mantle production ratios, and assuming initial nebular neon isotopic compositions, results in the ^{20}Ne/^{22}Ne ratio changing from 13.360 to 13.357 as the nebular ^{21}Ne/^{22}Ne ratio evolves to a value similar to the present-day Galápagos mantle (^{21}Ne/^{22}Ne = 0.0336)[20]; that is, the ^{20}Ne/^{22}Ne changes in the third decimal place, which is not significant given our current analytical uncertainties. Likewise, evolution of a nebular ^{21}Ne/^{22}Ne ratio to a value of 0.0368 (Discovery sample EW9309_5D) would have produced an insignificant shift in the ^{20}Ne/^{22}Ne ratio to 13.35.

If we assume that the plume mantle had an initial ^{20}Ne/^{22}Ne ratio of 13.03 and a corresponding ^{21}Ne/^{22}Ne ratio of 0.0328, then as the ^{21}Ne/^{22}Ne ratio evolves to the observed Discovery sample EW9309_5D value of 0.0368 the ^{20}Ne/^{22}Ne ratio would decrease by 0.01 to a value of 13.02. This shift is less than the reported analytical uncertainty of 0.04 for the sample. For an initial ^{20}Ne/^{22}Ne of 13.03, the ^{20}Ne/^{22}Ne ratio would decrease by 0.08 to a value of 12.95 as the ^{21}Ne/^{22}Ne ratio evolved to a value similar to that of the depleted MORB mantle (^{21}Ne/^{22}Ne = 0.0578)[19]. This evolved ^{20}Ne/^{22}Ne ratio of 12.95 is still significantly higher than the MORB mantle ^{20}Ne/^{22}Ne ratio of 12.5 (see main text) and indicates that the plume and MORB mantle ^{20}Ne/^{22}Ne ratios cannot be related simply through nucleogenic ingrowth. Rather, the two reservoirs must have acquired neon from two distinct accretionary sources. We note that nucleogenic production of neon will serve only to decrease the ^{20}Ne/^{22}Ne ratios over time. Therefore, our conclusions about nebular neon in the deep mantle are robust. Given that the nucleogenic corrections for ^{20}Ne/^{22}Ne are smaller than or comparable to our analytical uncertainties, we have not made any corrections to our measured ^{20}Ne/^{22}Ne ratios.

Determining mantle ^{20}Ne/^{22}Ne–^{36}Ar/^{22}Ne–^{130}Xe/^{22}Ne ratios. Mantle ^{20}Ne/^{22}Ne ratios for Galápagos, Iceland, and Rochambeau were computed by extrapolating their ^{20}Ne/^{22}Ne–^{21}Ne/^{22}Ne systematics to intercept the plume-MORB mixing line defined above. The ^{36}Ar/^{22}Ne and ^{130}Xe/^{22}Ne ratios for Iceland and Rochambeau

© 2019 Springer Nature Limited. All rights reserved.

were then determined through extrapolation of the $^{20}Ne/^{22}Ne$–$^{36}Ar/^{22}Ne$ and $^{20}Ne/^{22}Ne$–$^{130}Xe/^{22}Ne$ systematics to these computed $^{20}Ne/^{22}Ne$ ratios. For MORB, the $^{20}Ne/^{22}Ne$ ratio was taken from ref. [19], and the $^{36}Ar/^{22}Ne$/ and $^{130}Xe/^{22}Ne$ were computed from measurements conducted by ref. [40]. The $^{36}Ar/^{22}Ne$ and $^{130}Xe/^{22}Ne$ ratios for CI chondrites were computed by dividing the average ^{36}Ar and ^{130}Xe abundances determined by ref. [17] by the average ^{22}Ne abundances. The $^{36}Ar/^{22}Ne$ and $^{130}Xe/^{22}Ne$ ratios for solar nebula gas were determined following the methods of refs [15,58]. Deep ocean water values are from ref. [59] while serpentine and its metamorphosed equivalents are from refs [23,60,61]. All values are reported in Extended Data Table 2.

Pre-compaction exposure and the neon isotopic composition of meteoritic materials. Models for the origin of volatiles from solar wind implantation into early Solar System dust can be tested by investigating the pre-compaction exposure of chondrules. Pre-compaction exposure refers to the period of time when dust grains (chondrule precursors) are free-floating objects transported to regions of the gaseous protoplanetary disk where they are irradiated by either solar or galactic radiation before accretion onto their respective parent bodies. Several studies have used the neon isotopic composition of individual chondrules to investigate their pre-compaction history with regards to solar wind irradiation[62–72]. The vast majority of chondrules show no evidence for solar wind irradiation. The few chondrules that do show such evidence extrapolate to initial $^{20}Ne/^{22}Ne$ ratios of 12.5 or less, values that are much lower than the maximum measured $^{20}Ne/^{22}Ne$ ratios of plume-influenced basalts. Furthermore, the few chondrules that show excess exposure to solar wind are associated with matrix materials that also show excesses in neon isotopic composition (for example ref. [69]). These chondrule–matrix associations are only found in regolith breccias where the implantation of solar wind occurred during regolith formation after their parent bodies had accreted and after the nebular gas had dissipated. The association of regolith breccias with implanted solar wind also applies to the enstatite chondrites. Relatively high $^{20}Ne/^{22}Ne$ ratios have been reported for enstatite chondrites known to be regolith breccias. However, when excluding regolith breccias, the $^{20}Ne/^{22}Ne$ ratio of enstatite chondrites are all less than 12 (refs [73–75]). Similarly, aubrites, which are thought to be the differentiated products of enstatite chondrites, display $^{20}Ne/^{22}Ne$ ratios less than 12.1 ± 0.2 (refs [76,77]). Given this, enstatite chondrites, their differentiated products and other solar-wind-implanted materials cannot be responsible for the high $^{20}Ne/^{22}Ne$ ratios observed in the plume mantle.

Data availability

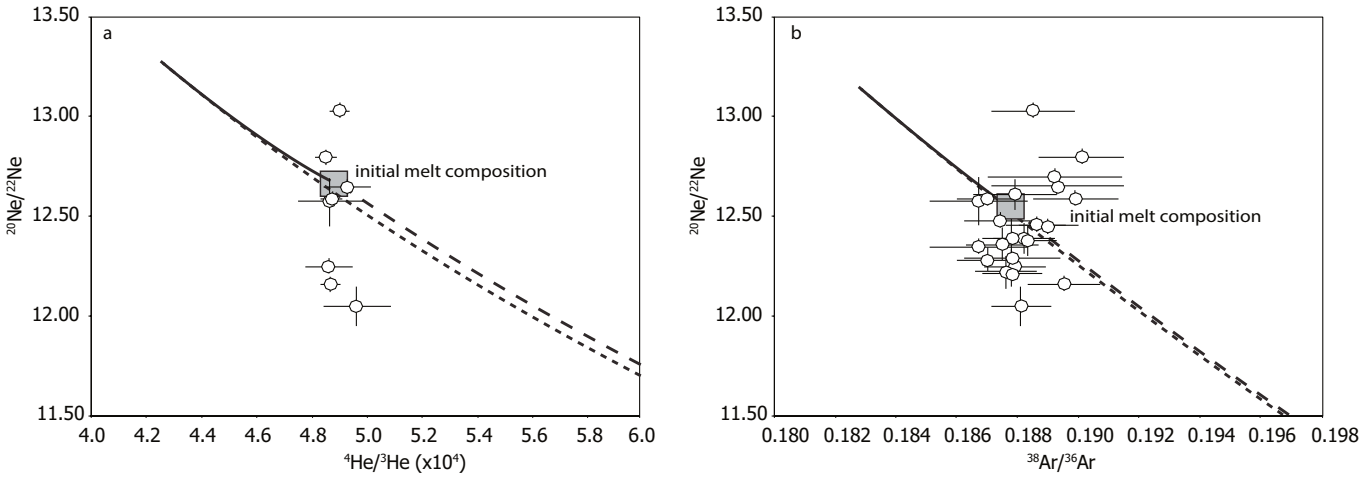
The main data supporting the findings of this study are available within the article, its Extended Data and Supplementary Table 1, as well as in the EarthChem database (https://doi.org/10.1594/IEDA/111217).

32. Gayer, E., Mukhopadhyay, S. & Meade, B. J. Spatial variability of erosion rates inferred from the frequency distribution of cosmogenic ^{3}He in olivines from Hawaiian river sediments. *Earth Planet. Sci. Lett.* **266**, 303–315 (2008).
33. Taylor, J. *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements* 2nd edn (University Science Books, Sausalito, 1997).
34. Courtillot, V., Davaille, A., Besse, J. & Stock, J. Three distinct types of hotspots in the Earth's mantle. *Earth Planet. Sci. Lett.* **205**, 295–308 (2003).
35. Gale, A., Dalton, C. A., Langmuir, C. H., Su, Y. & Schilling, J. G. The mean composition of ocean ridge basalts. *Geochem. Geophys. Geosyst.* **14**, 489–518 (2013).
36. Hanyu, T. et al. Noble gas study of the Reunion hotspot: evidence for distinct less-degassed mantle sources. *Earth Planet. Sci. Lett.* **193**, 83–98 (2001).
37. Honda, M. & Woodhead, J. D. A primordial solar-neon enriched component in the source of EM-I-type ocean island basalts from the Pitcairn Seamounts, Polynesia. *Earth Planet. Sci. Lett.* **236**, 597–612 (2005).
38. Moreira, M., Kunz, J. & Allegre, C. Rare gas systematics in popping rock: isotopic and elemental compositions in the upper mantle. *Science* **279**, 1178–1181 (1998).
39. Moreira, M. & Allègre, C. J. Rare gas systematics on Mid Atlantic Ridge (37–40° N). *Earth Planet. Sci. Lett.* **198**, 401–416 (2002).
40. Parai, R., Mukhopadhyay, S. & Lassiter, J. C. New constraints on the HIMU mantle from neon and helium isotopic compositions of basalts from the Cook–Austral Islands. *Earth Planet. Sci. Lett.* **277**, 253–261 (2009).
41. Parai, R., Mukhopadhyay, S. & Standish, J. J. Heterogeneous upper mantle Ne, Ar and Xe isotopic compositions and a possible Dupal noble gas signature recorded in basalts from the Southwest Indian Ridge. *Earth Planet. Sci. Lett.* **359**, 227–239 (2012).
42. Pető, M. K., Mukhopadhyay, S. & Kelley, K. A. Heterogeneities from the first 100 million years recorded in deep mantle noble gases from the Northern Lau Back-arc Basin. *Earth Planet. Sci. Lett.* **369–370**, 13–23 (2013).
43. Pető, M. K. *Application of Noble Gas Isotopic Systems to Identify Mantle Heterogeneities*. PhD thesis, Harvard Univ. (2014).
44. Stroncik, N. A. & Niedermann, S. He, Ne and Ar isotope signatures of mid-ocean ridge basalts and their implications for upper mantle structure: a case study from the Mid-Atlantic Ridge at 4–12° S. *Geochim. Cosmochim. Acta* **183**, 94–105 (2016).
45. Tucker, J. M., Mukhopadhyay, S. & Schilling, J. G. The heavy noble gas composition of the depleted MORB mantle (DMM) and its implications for the preservation of heterogeneities in the mantle. *Earth Planet. Sci. Lett.* **355–356**, 244–254 (2012).
46. Valbracht, P. J., Staudacher, T., Malahoff, A. & Allègre, C. J. Noble gas systematics of deep rift zone glasses from Loihi Seamount, Hawaii. *Earth Planet. Sci. Lett.* **150**, 399–411 (1997).
47. Moreira, M., Staudacher, T., Sarda, P., Schilling, J. G. & Allègre, C. J. A primitive plume neon component in MORB: the Shona ridge-anomaly, South Atlantic (51–52° S). *Earth Planet. Sci. Lett.* **133**, 367–377 (1995).
48. Moreira, M., Valbracht, P. J., Staudacher, T. & Allègre, C. J. Rare gas systematics in Red Sea Ridge basalts. *Geophys. Res. Lett.* **23**, 2453–2456 (1996).
49. Raquin, A. & Moreira, M. Atmospheric $^{38}Ar/^{36}Ar$ in the mantle: implications for the nature of the terrestrial parent bodies. *Earth Planet. Sci. Lett.* **287**, 551–558 (2009).
50. Moreira, M. et al. Rare gas systematics on Lucky Strike basalts (37° N, North Atlantic): evidence for efficient homogenization in a long-lived magma chamber system? *Geophys. Res. Lett.* **38**, L08304 (2011).
51. Ballentine, C. J. & Burnard, P. G. Production, release and transport of noble gases in the continental crust. *Rev. Mineral. Geochem.* **47**, 481–538 (2002).
52. Leya, I. & Wieler, R. Nucleogenic production of Ne isotopes in Earth's crust and upper mantle induced by alpha particles from the decay of U and Th. *J. Geophys. Res. Solid Earth* **104**, 15439–15450 (1999).
53. Yatsevich, I. & Honda, M. Production of nucleogenic neon in the Earth from natural radioactive decay. *J. Geophys. Res. Solid Earth* **102**, 10291–10298 (1997).
54. Hünemohr, H. *Edelgase in U- und Th-reichen Mineralen und die Bestimmung der 21Ne-Dicktarget-Ausbeute der 18O(α,n)21Ne-Kernreaktion im Bereich 4.0–8.8 MeV*. PhD thesis, Johannes-Gutenberg Univ. (1989).
55. Kyser, T. K. & Rison, W. Systematics of rare gas isotopes in basic lavas and ultramafic xenoliths. *J. Geophys. Res. Solid Earth* **87**, 5611–5630 (1982).
56. Kennedy, B. M., Hiyagon, H. & Reynolds, J. H. Crustal neon: a striking uniformity. *Earth Planet. Sci. Lett.* **98**, 277–286 (1990).
57. Ballentine, C. J. Resolving the mantle He/Ne and crustal $^{21}Ne/^{22}Ne$ in well gases. *Earth Planet. Sci. Lett.* **152**, 233–249 (1997).
58. Meshik, A., Hohenberg, C., Pravdivtseva, O. & Burnett, D. Heavy noble gases in solar wind delivered by Genesis mission. *Geochim. Cosmochim. Acta* **127**, 326–347 (2014).
59. Jenkins, W. J. et al. The deep distributions of helium isotopes, radiocarbon, and noble gases along the US GEOTRACES East Pacific Zonal Transect (GP16). *Mar. Chem.* **201**, 167–182 (2018).
60. Kendrick, M. A. et al. Subduction zone fluxes of halogens and noble gases in seafloor and forearc serpentinites. *Earth Planet. Sci. Lett.* **365**, 86–96 (2013).
61. Kendrick, M. A., Scambelluri, M., Hermann, J. & Padron-Navarta, J. A. Halogens and noble gases in serpentinites and secondary peridotites: implications for seawater subduction and the origin of mantle neon. *Geochim. Cosmochim. Acta* **235**, 285–304 (2018).
62. Beyersdorf-Kuis, U., Ott, U. & Trieloff, M. Early cosmic ray irradiation of chondrules and prolonged accretion of primitive meteorites. *Earth Planet. Sci. Lett.* **423**, 13–23 (2015).
63. Das, J. P. & Murty, S. V. S. Cosmogenic and trapped noble gases in individual chondrules: clues to chondrule formation. *Meteorit. Planet. Sci.* **44**, 1797–1818 (2009).
64. Das, J. P., Goswami, J. N., Pravdivtseva, O. V., Meshik, A. P. & Hohenberg, C. M. Cosmogenic neon in grains separated from individual chondrules: evidence of precompaction exposure in chondrules. *Meteorit. Planet. Sci.* **47**, 1869–1883 (2012).
65. Eugster, O., Lorenzetti, S., Krähenbühl, U. & Marti, K. Comparison of cosmic-ray exposure ages and trapped noble gases in chondrule and matrix samples of ordinary, enstatite, and carbonaceous chondrites. *Meteorit. Planet. Sci.* **42**, 1351–1371 (2007).
66. Nakamura, T., Nagao, K., Metzler, K. & Takaoka, N. Heterogeneous distribution of solar and cosmogenic noble gases in CM chondrites and implications for the formation of CM parent bodies. *Geochim. Cosmochim. Acta* **63**, 257–273 (1999).
67. Ott, U., Wieler, R. & Huber, L. Comment on "Cosmogenic neon in grains separated from individual chondrules: evidence of precompaction exposure in chondrules" by J. P. Das, J. N. Goswami, O. V. Pravdivtseva, A. P. Meshik, and C. M. Hohenberg. *Meteorit. Planet. Sci.* **48**, 1524–1528 (2013).
68. Polnau, E., Eugster, O., Burger, M., Krähenbühl, U. & Marti, K. Precompaction exposure of chondrules and implications. *Geochim. Cosmochim. Acta* **65**, 1849–1866 (2001).
69. Riebe, M. E. et al. Cosmogenic He and Ne in chondrules from clastic matrix and a lithic clast of Murchison: no pre-irradiation by the early sun. *Geochim. Cosmochim. Acta* **213**, 618–634 (2017).
70. Roth, A. S. G., Baur, H., Heber, V. S., Reusser, E. & Wieler, R. Cosmogenic helium and neon in individual chondrules from Allende and Murchison: implications for the precompaction exposure history of chondrules. *Meteorit. Planet. Sci.* **46**, 989–1006 (2011).
71. Roth, A. S., Metzler, K., Baumgartner, L. P. & Leya, I. Cosmic-ray exposure ages of chondrules. *Meteorit. Planet. Sci.* **51**, 1256–1267 (2016).
72. Roth, A. S., Metzler, K., Baumgartner, L. P., Hofmann, B. A. & Leya, I. Protracted storage of CR chondrules in a region of the disk transparent to galactic cosmic rays. *Meteorit. Planet. Sci.* **52**, 2166–2177 (2017).

© 2019 Springer Nature Limited. All rights reserved.

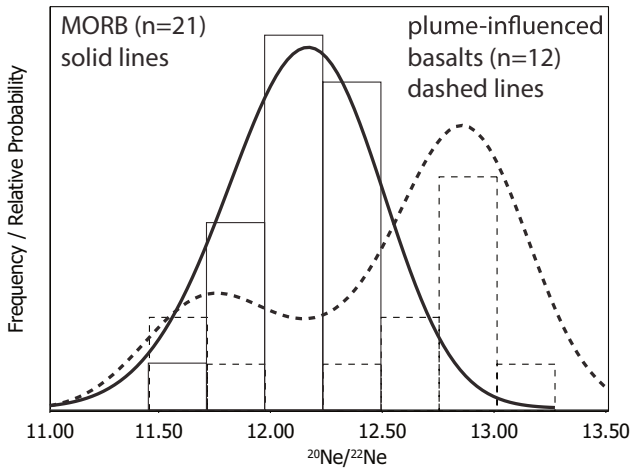
73. Okazaki, R., Takaoka, N., Nagao, K. & Nakamura, T. Noble gases in enstatite chondrites released by stepped crushing and heating. *Meteorit. Planet. Sci.* **45**, 339–360 (2010).
74. Wacker, J. F. & Marti, K. Noble gas components in clasts and separates of the Abee meteorite. *Earth Planet. Sci. Lett.* **62**, 147–158 (1983).
75. Wieler, R., Baur, H., Graf, T. & Signer, P. He, Ne, and Ar in Antarctic meteorites: solar noble gases in an enstatite chondrite. In *Proc. Lunar and Planetary Science Conference* Vol. 16, 903–904 (AGU, Washington DC, 1985).
76. Lorenzetti, S. et al. History and origin of aubrites. *Geochim. Cosmochim. Acta* **67**, 557–571 (2003).
77. Miura, Y. N., Hidaka, H., Nishiizumi, K. & Kusakabe, M. Noble gas and oxygen isotope studies of aubrites: a clue to origin and histories. *Geochim. Cosmochim. Acta* **71**, 251–270 (2007).
78. Rayleigh, L. L. Theoretical considerations respecting the separation of gases by diffusion and similar processes. *Lond. Edinb. Dublin Phil. Mag. J. Sci.* **42**, 493–498 (1896).

© 2019 Springer Nature Limited. All rights reserved.



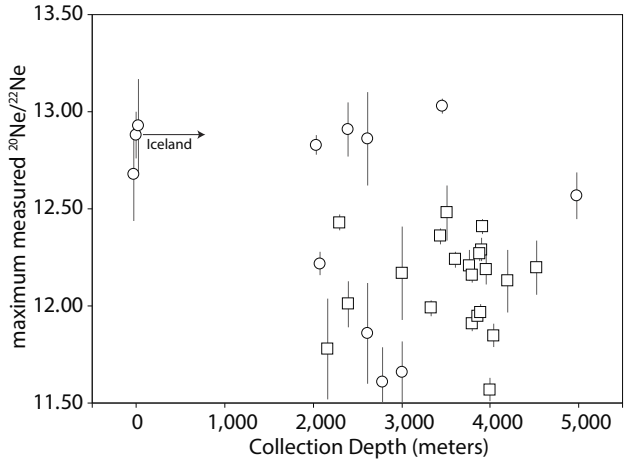
Extended Data Fig. 1 | Lack of mass-dependent isotope fractionation in the step-crushing neon data. Péron et al.[8] suggested that using the highest measured $^{20}Ne/^{22}Ne$ ratios for characterizing the plume mantle is inappropriate because mass-dependent isotope fractionation may occur during bubble formation[8]. In this scenario, mass-dependent fractionation during bubble formation would lead to $^{20}Ne/^{22}Ne$ ratios scattering about the 'true' mean value, with some bubbles characterized by relatively high $^{20}Ne/^{22}Ne$ ratios while other bubbles displayed relatively low $^{20}Ne/^{22}Ne$ ratios. This hypothesis can be tested by measuring $^{4}He/^{3}He$ and $^{38}Ar/^{36}Ar$ ratios during the same step-crushes as the neon isotopes as illustrated here. **a**, **b**, The measured neon isotopic compositions of individual step-crushes plotted against those of helium (**a**) and argon (**b**) along with predicted trajectories of mass-dependent isotope fractionation during bubble formation obtained by applying a Rayleigh fractionation model[78]. Here, the parental melt is assumed to have an initial $^{20}Ne/^{22}Ne$ ratio of 12.65 ± 0.08, similar to the value of ref. [8]. Initial helium and argon isotopic compositions are from the mean values determined for sample EW9309_5D in this study (Supplementary Table 1). The curves show the trajectory of the melts during degassing, the evolution of an instantaneously lost vapour phase (bubbles; short-dashed line) and the cumulative evolution of the vapour phase (bubbles; solid line). Plotted along with these curves are the individual step-crushes (circles) from this study (sample EW9309_5D) with their associated 2σ uncertainties (error bars). Note that the $^{4}He/^{3}He$ ratios were measured only on one aliquot of EW9309_5D. The helium–neon–argon isotopic compositions measured in the individual step-crushes do not follow the predicted Rayleigh fractionation trends. Rather, the data cloud is at a high angle to the predicted isotope fractionation trend. For example, for our highest measured $^{20}Ne/^{22}Ne$ of 13.03 ± 0.04 (2σ) to be a result of mass fractionation, the measured $^{38}Ar/^{36}Ar$ should be 0.1847 (**b**). However, the measured $^{38}Ar/^{36}Ar$ of 0.1885 ± 0.0014 (2σ) is identical to the atmospheric value and similar to other determinations of $^{38}Ar/^{36}Ar$ ratios in plumes and MORBs. Given this, we conclude that mass-dependent isotope fractionation during bubble formation is not responsible for generating the highest $^{20}Ne/^{22}Ne$ ratios determined in these studies.

© 2019 Springer Nature Limited. All rights reserved.



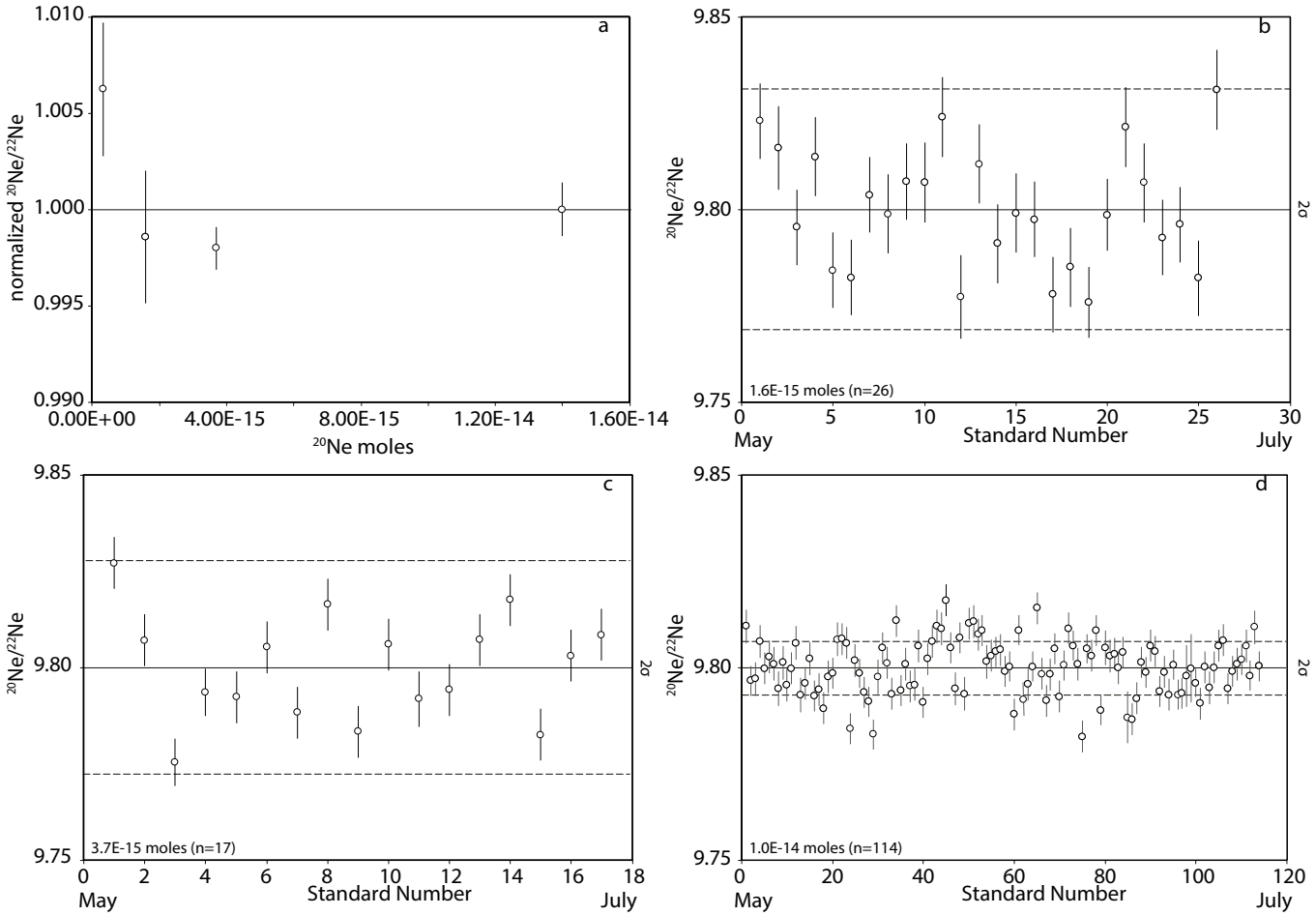
Extended Data Fig. 2 | Frequency functions for non-plume-influenced MORBs MORBs and plume-influenced basalts. Histograms and kernel density estimates were constructed from the maximum measured neon isotopic composition ($^{20}Ne/^{22}Ne_{max}$) of globally distributed, mantle-derived samples from non-plume-influenced MORBs (solid curve) and plume-influenced basalts (dashed curve), similar to Fig. 2. Kernel density estimates were calculated using the Matlab Curve Fitting Toolbox with bandwidths of 0.16 and 0.12 for non-plume-influenced MORBs and plume-influenced materials, respectively. The kernel density estimator results in slightly broader distributions than observed in Fig. 2 but does not change the main conclusions that there is a clear difference between the maximum measured $^{20}Ne/^{22}Ne$ ratios for non-plume-influenced MORBs and for plume-influenced basalts, with MORBs displaying a sharp cut-off at a $^{20}Ne/^{22}Ne$ ratio of 12.5.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | Collection depths of non-plume-influenced MORB samples and plume-influenced basalts. Depths are in metres. MORB samples (squares) highlighted in this study all erupted from depths between 2,000 and 5,000 m below sea level, and plume-influenced basalts (circles) on average erupted at comparable or shallower depths. For comparable eruption depths, plume-influenced basalts show higher $^{20}Ne/^{22}Ne$ ratios than non-plume-influenced MORBs. Moreover, if atmospheric contamination played a role in generating the difference between these two populations, plume-influenced basalts should have lower $^{20}Ne/^{22}Ne$ ratios, given their shallower eruption depth in the sample suite. However, such a relationship is not observed. Therefore, we conclude that different eruption depths are not responsible for the two distinct modes observed for non-plume-influenced MORBs and for plume-influenced basalts shown in Fig. 2. We note that the depth of eruption for Iceland sample DICE 10 is unknown, as it was erupted subglacially. Here, we have assigned a value of zero metres below sea-level to the DICE 10 samples, but deeper eruption depths will not change the results of this study. Data sources are reported in Extended Data Table 1.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | Long-term external reproducibility of bracketing standards. a, Mass discrimination of the $^{20}Ne/^{22}Ne$ ratio as a function of the ^{20}Ne beam size. The $^{20}Ne/^{22}Ne$ ratios for the different size standards were normalized to the $^{20}Ne/^{22}Ne$ ratio of the largest standard (10^{-14} moles of ^{20}Ne). The error bars (2σ uncertainties) reflect the relative errors in the $^{20}Ne/^{22}Ne$ isotope ratio based on the reproducibility of the standards. **b–d**, Reproducibility of standard $^{20}Ne/^{22}Ne$ ratios that were interspersed with the sample measurements over the 3-month period that it took to conduct all step-crushes. Error bars on the individual air standards represent the internal measurement error (2SE), while the dashed lines represent the long-term (2σ) external reproducibility. The external reproducibilities on the $^{20}Ne/^{22}Ne$ ratios were 0.03, 0.03 and 0.01 (2σ) for ^{20}Ne beam sizes of 1.6×10^{-15} moles ($n = 26$), 3.7×10^{-15} moles ($n = 17$) and 1×10^{-14} moles ($n = 114$), respectively.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Maximum measured $^{20}Ne/^{22}Ne$ and $^{21}Ne/^{22}Ne$ ratios for non-plume-influenced MORBs and plume-influenced materials

Sample	Ocean/Continent	Feature	Description	Crush or Heating	Collection Depth (meters)	$^{20}Ne/^{22}Ne$	2 sigma	$^{21}Ne/^{22}Ne$	2 sigma	References
MW87-12-69-4	Pacific	Plume-influenced materials	East Pacific Rise - 17S anomaly	Heating	2615	12.86	0.24	0.0407	0.0014	[21]
D22B	Pacific	Plume-influenced	Galapagos	Crush	2390	12.91	0.14	0.0336	0.0006	[20]
S2335-6	Pacific	Plume-influenced	Hawaii - Loihi	Heating	4983	12.57	0.12	0.0362	0.0010	[46]
Aliquot 5	Atlantic	Plume-influenced	Iceland - DICE 10	Crush	0	12.88	0.12	0.0355	0.0012	[14]
Magnetite	Asia	Plume-influenced	Kola Russia	Crush	0	12.93	0.24	0.0421	0.0042	[13]
NLD-27 I	Australian-Pacific	Plume-influenced	Lau Basin - Rochambeau Rift	Crush	2076	12.22	0.06	0.0430	0.0004	[42]
52DS1-a	Pacific	Plume-influenced	Pitcairn	Heating	3000	11.66	0.16	0.0324	0.0008	[37]
RF94-01 nodule	Indian	Plume-influenced	Reunion Piton de la Fournaise	Crush	0	12.68	0.24	0.0395	0.0018	[36]
23D-1g'	Atlantic	Plume-influenced	Shona	Heating	2609	11.86	0.26	0.0419	0.0022	[47]
MD73 C1012	Red Sea	Plume-influenced	Red Sea	Heating	2777	11.61	0.18	0.0426	0.0014	[48]
EW9309_5D	Atlantic	Plume-influenced	Discovery	Crush	3453	13.03	0.04	0.0368	0.0007	This study
EW9309_25D	Atlantic	Plume-influenced	Discovery	Crush	2032	12.83	0.05	0.0470	0.0012	This study
131GTV4 M64/1	Atlantic	Spreading center	MORBs N and S of Ascension	Heating	2999	12.17	0.24	0.0608	0.0050	[44]
AG22 1-4	Indian	Spreading center	Southwest Indian Ridge	Crush	4000	11.57	0.06	0.0552	0.0010	[41]
All 107-6 57-5	Indian	Spreading center	Southwest Indian Ridge	Crush	3767	12.21	0.08	0.0583	0.0006	[41]
CH98DR12	Atlantic	Spreading center	Mid-Atlantic Ridge	Heating	4200	12.13	0.16	0.0516	0.0020	[39]
D9, Hotu Matua Smt.	Pacific	Seamount	East Pacific Rise	Crush	3330	11.99	0.04	0.0548	0.0008	[40]
EN061 4D	Atlantic	Spreading center	Equatorial mid-Atlantic Ridge	Crush	2300	12.43	0.04	0.0636	0.0012	[45]
HLY0102-006-st gl.	Arctic	Spreading center	Gakkel Ridge - East Volcanic Zone	Crush	3800	11.91	0.04	0.0560	0.0006	[43]
HLY0102-008-001 repeat	Arctic	Spreading center	Gakkel Ridge - East Volcanic Zone	Crush	4042	11.85	0.06	0.0561	0.0010	[43]
HLY0102-026-020	Arctic	Spreading center	Gakkel Ridge - East Volcanic Zone	Crush	3856	11.95	0.02	0.0550	0.0008	[43]
HLY0102-062-st gl	Arctic	Spreading center	Gakkel Ridge - East Volcanic Zone	Crush	3900	12.29	0.06	0.0591	0.0012	[43]
HLY0102 D67-1 repeat	Arctic	Spreading center	Gakkel Ridge - East Volcanic Zone	Crush	3954	12.19	0.08	0.0590	0.0016	[43]
KN162-7 11-25	Indian	Spreading center	Southwest Indian Ridge	Crush	3913	12.41	0.04	0.0640	0.0014	[41]
KN162-7 18-17	Indian	Spreading center	Southwest Indian Ridge	Crush	4525	12.20	0.14	0.0598	0.0024	[41]
KN162-7 23-107	Indian	Spreading center	Southwest Indian Ridge	Crush	3609	12.24	0.04	0.0635	0.0010	[41]
RC2806 42D-7	Atlantic	Spreading center	Equatorial mid-Atlantic Ridge	Crush	3440	12.36	0.04	0.0569	0.0012	[45]
RC2806 48D-9	Atlantic	Spreading center	Equatorial mid-Atlantic Ridge	Crush	3896	11.97	0.04	0.0539	0.0008	[45]
RC2806 57D-1	Atlantic	Spreading center	Equatorial mid-Atlantic Ridge	Crush	3885	12.27	0.04	0.0580	0.0012	[45]
RC2806 3D-2	Atlantic	Spreading center	Equatorial mid-Atlantic Ridge	Crush	3800	12.16	0.04	0.0591	0.0010	[45]
2PiD43	Atlantic	Spreading center	Equatorial mid-Atlantic Ridge - Popping Rock	Crush	3510	12.48	0.14	0.0599	0.0016	[38]
SWIFT bis Sc05-1-1 (1)	Indian	Spreading center	Southwest Indian Ridge	Crush	2397	12.01	0.12	0.0534	0.0042	[49]
GRA RC07	Atlantic	Spreading center	Lucky Strike Region mid-Atlantic Ridge	Heating	2160	11.78	0.26	0.0491	0.0042	[50]

Data from this study and refs [13,14,20,21,36–50].

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Noble gas end-member compositions

End-member	$^{20}Ne/^{22}Ne$	2 sigma	$^{36}Ar/^{22}Ne$	2 sigma	$^{130}Xe/^{22}Ne$	2 sigma	References
Galapagos	13.23	0.22	n.d.	n.d.	n.d.	n.d.	[20]
Iceland	13.17	0.22	4.61	1.00	0.0029	0.0005	[14]
Rochambeau	12.85	0.25	10.83	2.20	0.0081	0.0019	[42]
MORB	12.49	0.08	11.60	11.80	0.0099	0.0056	[19]
CI chondrites	9.03	2.46	19.50	16.00	0.0343	0.0368	[17]
Solar nebula gas	13.36	0.18	0.33	range=0.33 to 0.40	0.0002	range=0.0002 to 0.0003	[15]
Solar wind	13.78	0.02	0.33	0.02	2.30E-06	2.00E-07	[15, 58]
Deep seawater (>2,500m)	9.80	-	73.66	1.08	0.0330	0.0020	[59]
Serpentinite and its metamorphosed equivalents	9.80	-	68.00	75.20	0.1487	range=0.016 to 0.499	[23, 60-61]
Air	9.80	-	18.72	-	0.0021	-	

n.d. not determined.

Data from refs [14,15,17,19,20,23,42,58–61].

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0710-1

Late Middle Pleistocene Levallois stone-tool technology in southwest China

Yue Hu[1], Ben Marwick[1,2]*, Jia-Fu Zhang[3], Xue Rui[1], Ya-Mei Hou[4,5], Jian-Ping Yue[4,5], Wen-Rong Chen[6], Wei-Wen Huang[4] & Bo Li[1,7]*

Levallois approaches are one of the best known variants of prepared-core technologies, and are an important hallmark of stone technologies developed around 300,000 years ago in Africa and west Eurasia[1,2]. Existing archaeological evidence suggests that the stone technology of east Asian hominins lacked a Levallois component during the late Middle Pleistocene epoch and it is not until the Late Pleistocene (around 40,000–30,000 years ago) that this technology spread into east Asia in association with a dispersal of modern humans. Here we present evidence of Levallois technology from the lithic assemblage of the Guanyindong Cave site in southwest China, dated to approximately 170,000–80,000 years ago. To our knowledge, this is the earliest evidence of Levallois technology in east Asia. Our findings thus challenge the existing model of the origin and spread of Levallois technologies in east Asia and its links to a Late Pleistocene dispersal of modern humans.

Middle Palaeolithic prepared-core reduction strategies, commonly referred to as Levallois or mode III technologies, are remarkable for both their ubiquity in Eurasia and Africa, and their apparent absence in east Asia during the Middle and Late Pleistocene (Fig. 1). This uneven global distribution has obscured the origins of Levallois technology, and its relationship to later technologies. The appearance of Levallois artefacts around 200–300 thousand years ago (ka) in Eurasia marks the transition from the Lower to Middle Palaeolithic in these regions. This was a major innovation in optimizing lithic tool manufacturing[3], potentially signalling an expansion of archaic *Homo* populations from Africa[4]. However, early Levallois technology found with bifaces in the southern Caucasus suggests that Levallois technology evolved from the existing local Acheulian (or mode II) technological systems[5]. This supports a hypothesis of isolated technological convergence[6], rather than a single-origin and dispersal model. The recent discovery of Levallois technology in India from around 385–172 ka[7] also raised the need to re-evaluate the relationship between the origins of Middle Palaeolithic culture in South Asia and the dispersal of modern humans.

Previous archaeological evidence from China, Mongolia, South Korea and Japan[8–14] suggests that major changes in raw material procurement, core reduction, retouch and typology of stone artefacts in east Asia tend to be clustered at the Upper Pleistocene (Fig. 1), indicating that a distinct Middle Palaeolithic period of systematic technological innovation did not occur in eastern Asia[15]. Without early ancestral technologies such as the Levallois technology, the appearance of blades in the Upper Pleistocene in East Asia indicates that they may have resulted from population admixture or replacement. The apparent absence of the Levallois technology in east Asia similarly raises critical questions about the relationship between cultural and biological trajectories of populations in east Asia and western regions.

Here we describe the stone artefact assemblage from Guanyindong Cave in the Guizhou province, southwest China (Fig. 2a) that provide evidence of an early appearance of Levallois artefacts in East Asia. Discovered in 1964, Guanyindong Cave is a limestone cave (Extended Data Fig. 1). Excavations during 1964–1973 recovered more than 3,000 stone artefacts and numerous fossilized fauna[16]. Faunal remains mostly belong to the Middle Pleistocene *Ailuropoda–Stegodon* fauna complex (Supplementary Information). Several trenches were opened within and in front of the cave, but most of the artefacts were excavated from the main entrance located at the west end of the cave. The stratigraphy of the main entrance was divided into nine layers that can be attributed to three groups (groups A, B and C) (Extended Data Fig. 2 and Supplementary Information). Stone artefacts and fossils were found in groups A (layer 2) and B (layers 3–8) only. Because this site was excavated more than 40 years ago, only 204 pieces of the studied stone artefacts have clear stratigraphic information (87 artefacts from group A and 117 from group B). Among these, we identified five artefacts as Levallois; three of these (two cores and one tool) are from group A and two (all tools) are from group B (Extended Data Figs. 5–7, 10). This suggests that Levallois concepts were present at this site throughout the whole occupation period.

This site was previously dated by U-series techniques[17,18], and a wide range of U-series ages ranging from around 50 to about 240 thousand years (kyr) old have been reported (Supplementary Table 1). However, many of these U-series ages were made on fossils, which should be treated as minimum age estimates[19]. Furthermore, most of the dated carbonate samples do not have firm stratigraphic control, so it is unreliable to associate their U-series ages to the sediment layers (see Supplementary Information for a full discussion on U-series results). To confirm the age of the Guanyindong assemblage, we used single-grain optically stimulated luminescence (OSL) dating on quartz (see Methods) to determine the ages of the deposits from layer 1, groups A and B (Fig. 2b and Extended Data Figs. 3, 4). Three samples from layer 1 yielded age estimates of ~70–40 kyr. Four samples from group A yielded ages of around 90–80 kyr and six samples from group B yielded ages of around 170–160 kyr (Fig. 2b). The OSL ages obtained for each of the groups are statistically consistent with each other at 2σ. Our dating results suggest that both groups A and B were deposited over short periods, although there is a large gap in age (around 80 kyr) between groups A and B, which is consistent with the observation of a sedimentary unconformity between the two groups (Extended Data Fig. 2). Our OSL chronology, therefore, securely places the date of deposit for the Guanyindong archaeological deposits (layers 2–8) between approximately 170 and about 80 ka.

The Guanyindong Cave assemblage consists of flakes, flake breaks, retouched pieces, cores, chunks and debris. The raw materials are predominantly chert (Extended Data Figs. 8–10; see Supplementary Information and Supplementary Figs. 21–24). On the basis of the detailed analysis of 2,273 stone artefacts, we found evidence of Levallois concepts in 45 specimens (see Methods and Supplementary Information for detailed justification), including 11 cores, 30 flakes and

[1]Centre for Archaeological Science, School of Earth and Environmental Sciences, University of Wollongong, Wollongong, New South Wales, Australia. [2]Department of Anthropology, University of Washington, Seattle, WA, USA. [3]MOE Laboratory for Earth Surface Processes, Department of Geography, College of Urban and Environmental Sciences, Peking University, Beijing, China. [4]Key Laboratory of Vertebrate Evolution and Human Origins, Institute of Vertebrate Paleontology and Paleoanthropology, Chinese Academy of Sciences, Beijing, China. [5]CAS Centre for Excellence in Life and Paleo-environment, Beijing, China. [6]Qianxi County Bureau of Cultural Relics Protection, Bijie, China. [7]ARC Centre of Excellence for Australian Biodiversity and Heritage, University of Wollongong, Wollongong, New South Wales, Australia. *e-mail: bmarwick@uw.edu; bli@uow.edu.au

© 2019 Springer Nature Limited. All rights reserved.

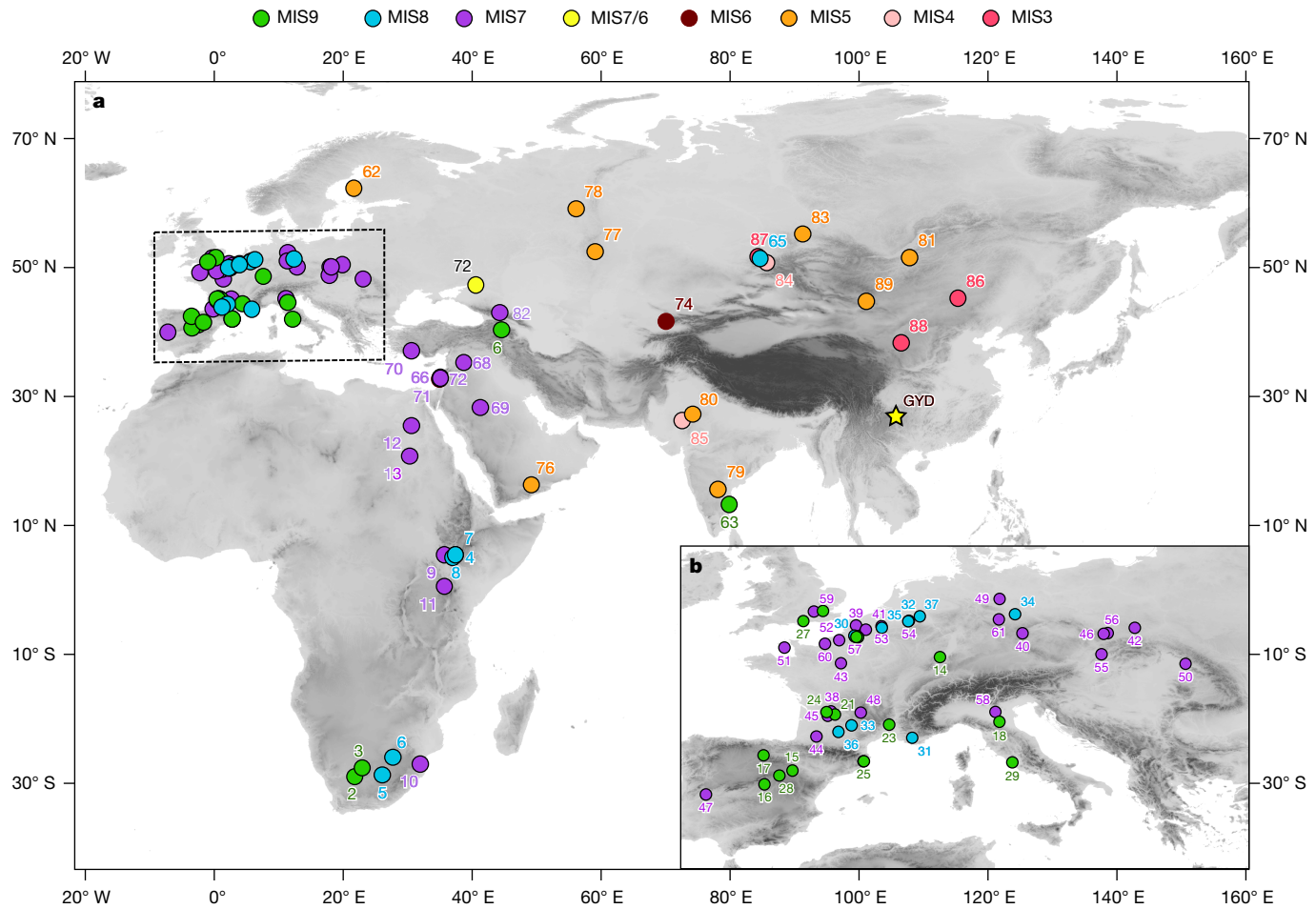


Fig. 1 | Distribution of Levallois technology during the late Middle Pleistocene (from MIS 9 to 3) in Africa and Eurasia. a, **b**, Distribution of Levallois technology across Africa and Eurasia. **b**, Magnification of the region inside the dashed rectangle in **a**. Detailed information on the sites is provided in Supplementary Table 2. The MIS corresponding to the chronology of individual sites is indicated by different colour-coded symbols. Note that there are a large number of sites that are younger than MIS 7 in Europe and Africa; however these sites are not shown here. GYD, Guanyindong Cave.

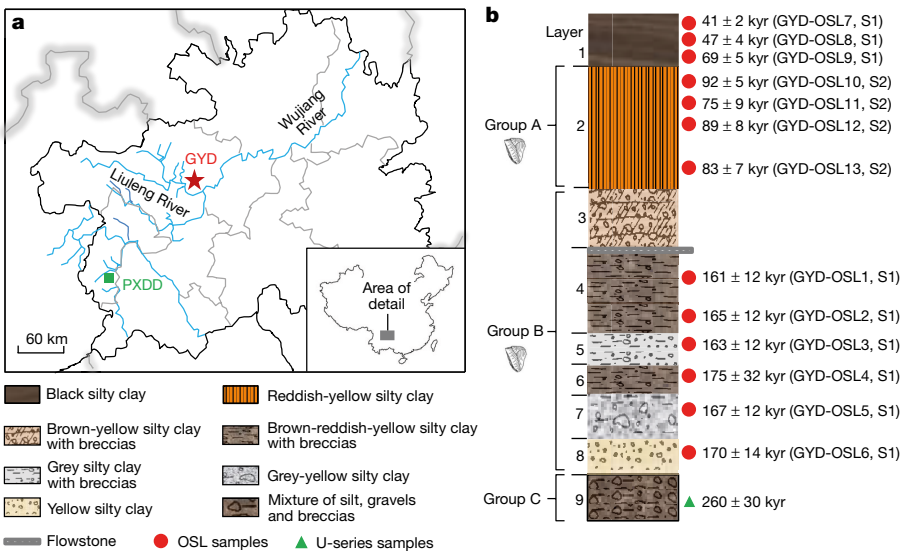


Fig. 2 | Location, stratigraphy and chronology of the Guanyindong site. a, Location of the Guanyindong Cave (GYD) and Panxian Dadong (PXDD) sites in Guizhou Province, southwest China. **b**, Schematic composite stratigraphy at the south wall of the cave entrance, with the depth, profile and ages of the OSL samples and U-series[17] dating results indicated. The sketches of stone tools indicate cultural layers. The uncertainties of the OSL ages are expressed at 1σ. S1 and S2 represent the two residual profiles at the south wall of the cave entrance (Extended Data Fig. 2) where the OSL samples were taken.

© 2019 Springer Nature Limited. All rights reserved.

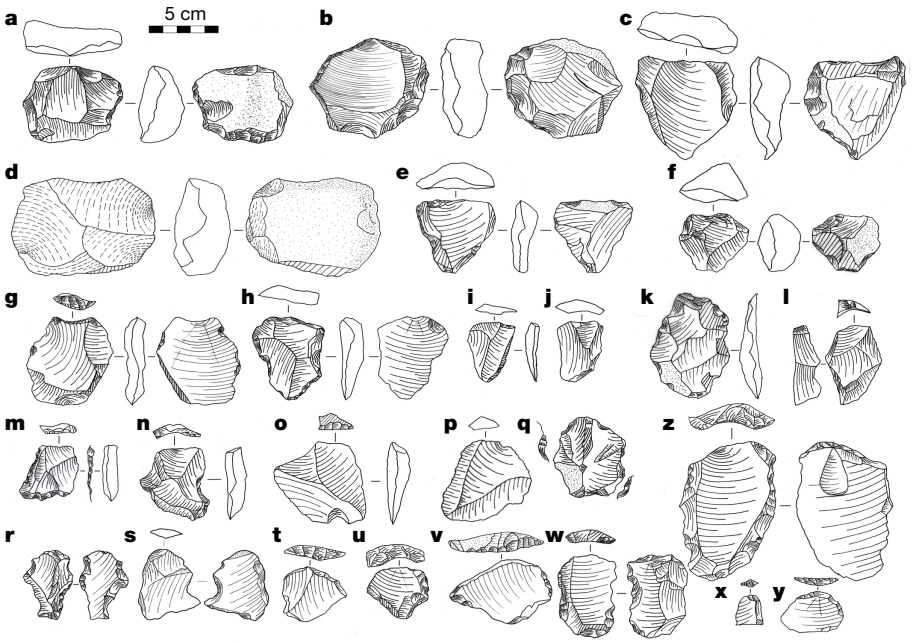


Fig. 3 | Line drawings of selected artefacts from Guanyindong Cave. **a**, **d**, **f**, Levallois recurrent cores. **b**, **c**, **e**, Levallois preferential cores. **g**–**k**, **n**, Levallois flakes. **l**, Débordant. **o**, **p**, Pseudo-Levallois points. **m**, **q**–**s**, Tools made on Levallois blanks. **t**–**z**, Flakes with prepared platforms. The photos of these artefacts are shown in Extended Data Figs. 5–7. The 3D structures of **a**–**c** are shown in the Supplementary Data. The artefacts shown in **b**, **c** and **q** were recovered from group A, and those shown in **r** and **s** were from group B.

4 tools made on Levallois flakes (Fig. 3 and Extended Data Figs. 5–7). Our technological reading of artefacts differs from previous studies, and in support of our analysis we provide three-dimensional models of three Levallois cores (shown in Fig. 3a–c) in the Supplementary Data. Eight cores exhibit patterns of recurrent Levallois concepts (Fig. 3a, d, f), each with two intersecting hierarchically organized surfaces. The upper surfaces of these cores are covered with several scars removed to form convexities that influence the pattern of detachment of the final flake. These scars come from different directions forming a centripetal scar pattern. The scars of the predetermined flakes are parallel to the plane of the intersection of upper and lower surface. The debitage surfaces of the cores have small flake scars along the edge, indicating preparation of their striking platforms. Three preferential Levallois cores are present (Fig. 3b, c, e), and are identifiable by the prominent large final flake detachments that have truncated the distal regions of the previous preparatory flake scars. The scars of the main flake removal on these cores are also parallel to the intersection of the upper and lower surfaces. The lower surfaces are extensively scarred and small platform preparation flake removals are present on the core circumference.

Many Levallois flakes at Guanyindong Cave exhibit a facetted platform, which results from core preparation before flake detachment. In addition, several smaller scars coming on to the dorsal surface of a flake from different directions are visible (Fig. 3g–k, n). These smaller scars may result from flaking to maintain the convexity of the core and in preparation for the removal of the Levallois flake. Four Levallois flakes were retouched along the edges (Fig. 3m, q–s). Besides these distinctive Levallois pieces, a number of non-Levallois flakes show signs of platform preparation (Fig. 3t–z), supporting the presence of more generalized strategies of prepared-core technology in Guanyindong Cave.

Levallois concepts at Guanyindong Cave first appeared in group B, which was dated to Marine Isotope Stage (MIS) 6 (approximately 180–130 ka), a period contemporary to the period during which Levallois technology was widely adopted in Africa and Eurasia[1]. Syntheses of globally distributed benthic $\delta^{18}O$ records indicate that MIS 6 was a glacial period of cooler temperatures and lower sea levels than at present[20]. Microscopic freeze–thaw features in the MIS 6 sediments from the nearby site Panxian Dadong (Fig. 2) suggest frequent freezing conditions during glacial periods, and the winter temperatures of this region during MIS 6 reached −5 °C or lower[21,22]. This evidence, together with the composition of the Panxian Dadong faunal assemblage, indicate a mixed woodland environment, including bamboo forests and open rocky areas with abundant grasses[21,22], suggesting that the landscape around Guanyindong Cave probably contained a reduced rainforest area compared to the present landscape, and a much-expanded open woodland environment.

The earliest age of the Guanyindong Cave lithic assemblage postdates the earliest modern human fossils in Africa[6,23] by 300–200 kyr and the Levant[24] by around 177–194 kyr, but predates any existing evidence of modern humans beyond this region during MIS 5 (around 130–80 ka), especially in south and southwest China[25,26]. With a secure age of approximately 170–80 kyr, the Levallois artefacts from Guanyindong Cave provide, to our knowledge, the earliest unequivocal evidence of prepared-core technology in east Asia, suggesting a geographically more widespread distribution of Levallois before the dispersal of *Homo sapiens*. This discovery has two important implications. First, the Guanyindong Cave assemblage suggests that demographic events may have occurred earlier in the Middle Pleistocene, leading to the appearance of Levallois concepts in east Asia. This possibility is suggested by the approximately 100 kyr-old Xuchang crania with its mosaic of Eurasian and Neanderthal features that indicate population interactions across Eurasia[27]. A Middle Pleistocene demographic event is also indicated by ancient DNA from the Late Pleistocene Tianyuan individual[28] that suggests that the divergence of Asians from Europeans occurred before 40 ka. Second, the emerging evidence of mode II bifacial tools from archaeological sites in east Asia[29,30] indicates that the prepared-core technologies from Guanyindong Cave, although rare, may alternatively represent a convergent technological evolution within the Acheulean technology of the same region. This challenges the existing hypotheses for the absence of Middle Pleistocene prepared-core technology in east Asia, including the idea that there was a lack of a strong ancestral Acheulean (mode II) tradition in this region and that local raw stone materials constrained tool-making to simple forms.

Given the absence of human fossils dated to the same period in southwest China, we can only speculate which species of hominin produced the Guanyindong Cave assemblage. Our findings, however, demonstrate a behavioural capacity compatible with their counterparts from the Western Hemisphere. The rarity of material traces of these complex behaviours in east Asia, relative to the Old World, therefore,

© 2019 Springer Nature Limited. All rights reserved.

may instead be due to the small, low-density populations with weak and/or irregular patterns of social interconnectedness in this region during the Middle Pleistocene. Under these conditions, technological innovation, transmission and persistence would have been rarer, compared to the high population and/or high density conditions of Middle Pleistocene sub-Saharan Africa, where Levallois is more abundant. Because Guanyindong Cave is one of only a few Palaeolithic sites that have been discovered in south China that are reliably dated to the late Middle Pleistocene, the abundance of mode III technology in this region remains an open question.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0710-1.

Received: 13 December 2017; Accepted: 26 September 2018;
Published online 19 November 2018.

1. Monnier, G. F. The Lower/Middle Paleolithic periodization in western Europe: an evaluation. *Curr. Anthropol.* **47**, 709–744 (2006).
2. Mcbrearty, S. & Brooks, A. S. The revolution that wasn't: a new interpretation of the origin of modern human behavior. *J. Hum. Evol.* **39**, 453–563 (2000).
3. Lycett, S. J. & Eren, M. I. Levallois lessons: the challenge of integrating mathematical models, quantitative experiments and the archaeological record. *World Archaeol.* **45**, 519–538 (2013).
4. Foley, R. & Lahr, M. M. Mode 3 technologies and the evolution of modern humans. *Camb. Archaeol. J.* **7**, 3–36 (1997).
5. Adler, D. S. et al. Early Levallois technology and the Lower to Middle Paleolithic transition in the Southern Caucasus. *Science* **345**, 1609–1613 (2014).
6. Tryon, C. A., McBrearty, S. & Texier, P.-J. Levallois lithic technology from the Kapthurin formation, Kenya: Acheulian origin and Middle Stone Age diversity. *Afr. Archaeol. Rev.* **22**, 199–229 (2005).
7. Akhilesh, K. et al. Early Middle Palaeolithic culture in India around 385–172 ka reframes Out of Africa models. *Nature* **554**, 97–101 (2018).
8. Boëda, E., Hou, Y. M., Forestier, H., Sarel, J. & Wang, H. M. Levallois and non-Levallois blade production at Shuidonggou in Ningxia, North China. *Quat. Int.* **295**, 191–203 (2013).
9. Li, F., Chen, F., Wang, Y. & Gao, X. Technology diffusion and population migration reflected in blade technologies in northern China in the Late Pleistocene. *Sci. China Earth Sci.* **59**, 1540–1553 (2016).
10. Brantingham, P. J., Olsen, J. W., Rech, J. A. & Krivoshapkin, A. I. Raw material quality and prepared core technologies in northeast Asia. *J. Archaeol. Sci.* **27**, 255–271 (2000).
11. Li, F. et al. The easternmost Middle Paleolithic (Mousterian) from Jinsitai Cave, North China. *J. Hum. Evol.* **114**, 76–84 (2018).
12. Seong, C. & Bae, C. J. The eastern Asian 'Middle Palaeolithic' revisited: a view from Korea. *Antiquity* **90**, 1151–1165 (2016).
13. Sato, H., Nishiaki, Y. & Suzuki, M. in *The Definition and Interpretation of Levallois Technology* (eds Dibble, H. L. & Bar-Yosef, O.) 485–500 (Prehistory Press, Madison, 1995).
14. Zwyns, N. in *Encyclopedia of Global Archaeology* (eds Gladyshev, S. et al.) 5025–5032 (Springer, New York, 2014).
15. Gao, X. & Norton, C. J. A critique of the Chinese 'Middle Palaeolithic'. *Antiquity* **76**, 397–412 (2002).
16. Li, Y. & Wen, B. *Guanyindong: A Lower Paleolithic Site at Qianxi County, Guizhou Province* (Cultural Relics Press, Beijing, China, 1986).
17. Shen, G. J. & Jin, L. H. U-series dating of speleothem samples from Guanyindong Cave at Qianxi County, Guizhou Province. *Acta Anthropologica Sinica* **11**, 93–100 (1992).
18. Yuan, S. X., Chen, T. M. & Gao, S. J. Uranium series chronological sequence of some Paleolithic sites in South China. *Acta Anthropologica Sinica* **5**, 179–190 (1986).
19. Grün, R., Eggins, S., Kinsley, L., Moseley, H., Sambridge, M. Laser ablation U-series analysis of fossil bones and teeth. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **416**, 150–167 (2014).
20. Lisiecki, L. E. & Raymo, M. E. A Pliocene–Pleistocene stack of 57 globally distributed benthic $\delta^{18}O$ records. *Paleoceanography* **20**, PA1003 (2005).
21. Wang, W. et al. Panxian Dadong, South China: establishing a record of Middle Pleistocene climatic changes. *Asian Perspect.* **43**, 302–313 (2004).
22. Karkanas, P., Schepartz, L. A., Miller-Antonio, S., Wang, W. & Huang, W. Late Middle Pleistocene climate in southwestern China: inferences from the stratigraphic record of Panxian Dadong Cave, Guizhou. *Quat. Sci. Rev.* **27**, 1555–1570 (2008).
23. Richter, D. et al. The age of the hominin fossils from Jebel Irhoud, Morocco, and the origins of the Middle Stone Age. *Nature* **546**, 293–296 (2017).
24. Hershkovitz, I. et al. The earliest modern humans outside Africa. *Science* **359**, 456–459 (2018).
25. Liu, W. et al. Human remains from Zhirendong, South China, and modern human emergence in East Asia. *Proc. Natl Acad. Sci. USA* **107**, 19201–19206 (2010).
26. Liu, W. et al. The earliest unequivocally modern humans in southern China. *Nature* **526**, 696–699 (2015).
27. Li, Z.-Y. et al. Late Pleistocene archaic human crania from Xuchang, China. *Science* **355**, 969–972 (2017).
28. Fu, Q. et al. DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl Acad. Sci. USA* **110**, 2223–2227 (2013).
29. Yamei, H. et al. Mid-Pleistocene Acheulean-like stone technology of the Bose basin, South China. *Science* **287**, 1622–1626 (2000).
30. Li, H., Kuman, K. & Li, C. What is currently (un)known about the Chinese Acheulean, with implications for hypotheses on the earlier dispersal of hominids. *C. R. Palevol.* **17**, 120–130 (2016).

Acknowledgements This work was supported by the Australian Research Council through Future Fellowships to B.L. (FT140100384) and B.M. (FT140100101), a grant from the National Science Foundation of China to J.-F.Z. (NSFC, 41471003), postgraduate scholarships from the University of Wollongong to Y.H. and X.R. and the China Scholarship Council to X.R. (201506010345), the Chinese Academy of Science (CAS) Strategic Priority Research Program Grants of 'Macroevolutionary Processes and Paleoenvironments of Major Historical Biota' (XDPB05), State Key Laboratory of Loess and Quaternary Geology, Institute of Earth Environment, CAS (SKLLQG1501) and National Science Foundation of China (41272033) to Y.-M.H. We thank S. Lin for assistance with artefact analysis and valuable comments on the manuscript; Y.-M. Hou for assistance with CT scanning on stone artefacts; R. G. Roberts, Z. Jacobs, Y. Jafari and T. Lachlan for support and assistance in the OSL laboratory; M. Otte and P. Zhang for valuable discussions on lithic assemblage; Y.-S. Lou, N. Ma, X.-W. Li and L. Lei for assistance with lithic observation.

Reviewer information *Nature* thanks C. A. Tryon and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions B.L., Y.H., W.-W.H. and J.-F.Z. conceived and coordinated the study; Y.H., B.M. and J.-P.Y. conducted the stone artefact analysis; B.L., Y.H., J.-F.Z., W.-R.C., W.-W.H. and Y.-M.H. planned and directed field investigations and stratigraphic analysis. Y.H., B.L., J.-F.Z., X.R., W.-W.H. and Y.-M.H. collected samples for dating; Y.H., B.L., J.-F.Z. and X.R. measured OSL samples and analysed the dating results; B.M., B.L. and Y.H. wrote the manuscript, with contributions from the other authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0710-1.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0710-1.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to B.M. or B.L.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

Artefact analysis. The concept of Levallois has a variety of definitions, so here we survey the variation in the use of this concept to establish how we identified artefacts as Levallois in the Guanyindong assemblage. At the centre of most modern definitions of Levallois technology are six technological criteria[31]: (1) exploitation of the volume of raw material is organized in terms of two intersecting planes or flaking surfaces; (2) the two surfaces are hierarchically related, one constituting the striking platform and the other the primary reduction surface; (3) the primary reduction surface is shaped such that the morphology of the product is predetermined, which is fundamentally a function of the lateral and distal convexities of the surface; (4) the fracture plane for removing primary products is sub-parallel to the plane of intersection of the two surfaces; (5) the striking platform size and shape is adjusted to allow removal of flakes parallel to this plane, usually through retouch or faceting; and (6) Levallois flakes are removed by direct hard hammer percussion.

This reduction sequence concept is the prevailing definition of Levallois technology worldwide. As noted previously[32–34], there are many possible core morphologies that are consistent with these six criteria. The specific actions required to achieve these criteria, such as cortex trimming, platform faceting and edge preparation, may be applied in different proportions and at different stages in the life of a core. Further variability is evident in patterns of surface preparation and the orientation of flake removals. Among this variability, three patterns of Levallois reduction have been documented, including flakes removed from along the circumference of the core (centripetal or radial), from two directions (orthogonal or opposed) or one from only direction (unidirectional, parallel or convergent). Within these patterns there are two basic systems: preferential, in which only one large flake is produced per core preparation episode and recurrent, where several large flakes are removed between each core preparation episode[31].

These variations in technical attributes may result in a wide range of shapes, but this does not alter the fundamental model of Levallois reduction. This technical approach to defining and identifying Levallois technology differs from the older Bordesian typological concept of the Levallois. The Bordesian definition is based on the presence of specific, visually distinctive core and flake products, such as the classic turtle-shell core and large detached central flake (that is, preferential Levallois flake) that are often depicted in explanations of Levallois technology[35,36]. A key point of contrast in the two definitions is that for the first, the distinctive innovation in Levallois technology is the result of a process or sequence of actions that produces cores with a distinctive geometry, whereas for the latter, the distinctive idea is the systematic production of artefacts with predetermined, visually distinctive shapes. Predetermination is also important in the first scheme; however, the visual distinctiveness and morphology of the product is less important. The broader implications are similar, that the artefact maker used foresight and planning to create a stone artefact. But the implications for identifying a Levallois assemblage are substantially different. The first concept permits many different flaking strategies within the Levallois and a wide diversity in the form and character of flake products[37]. On the other hand, if we use the more strict Levallois definition, we are constrained to forms that match the Mousterian typology and similarly precise and delicate pieces.

One distinctive technological strategy that is common to both definitions of Levallois is the preparation of the core platform between each flake removal. This is a key point that separates Levallois from discoidal reduction, where there is no intervening phase of remodelling the core between flake removals, and an unhierarchical relation of the surfaces (but see a previous study[38] for some of the debates surrounding discoids and Levallois). Traces of core platform preparation are also important for identifying foresight and planning in stone artefact production, which is the key behavioural implication for early evidence of Levallois. Core preparation for removal of a target flake is also the main concept of mode III technologies, of which the Levallois is the most intensively studied and best known subset. However, evidence of core preparation, although behaviourally important, is not by itself sufficient to identify Levallois technology in an assemblage. Similarly, the hierarchical organization of the surfaces by itself, without signs of preparation, is not sufficient to identify Levallois. For example, Middle Pleistocene hierarchical cores that do not show maintenance of distal and lateral convexities, and only minimal treatment of the preparatory surface is conducted, mainly by large removals, are not identified as Levallois[39,40]. Flakes resulting from these cores tend to be flat in terms of ventral curvature, with mostly plain striking platforms, showing no signs of platform preparation.

We see in previous work that when traces of core preparation are present, as well as some of the six criteria, but the overall artefact morphology is not typical of the Mousterian typology, that researchers hesitate to use the term 'Levallois'. Instead they use terms such as 'proto-Levallois', 'stripped-down Levallois'[41], 'Levallois-like'[10,42–44], 'unsophisticated Levallois'[45], 'para-Levallois'[46,47] or 'reduced Levallois'[48]. These terms are most common when discussing assemblages at the early chronological extreme of the European Middle Palaeolithic or African Middle Stone Age, or at geographical extremes of the classic Levallois area, such as China. In many cases this nomenclature reflects either transitional technologies from simple prepared cores to 'full' Levallois with core preparation and hierarchical surfaces[41] or localized, independent convergences on Levallois technology that have no historical connection to the Bordesian core area of Levallois[49], or simply are pieces that are less intensely modified, representing initial phases of knapping[50]. This raises the question: what are the limits of the Levallois definition?

A particularly problematic detail in establishing the limits of the definition is the means by which the hierarchical relationship between the two core surfaces was established and how the platform was prepared to orient it perpendicular to the axis of flaking. It has previously been noted[32] that the previously published definition[31] gives little guidance on this. Several studies identify cores with a morphology of naturally asymmetric surfaces as Levallois, even though they lack the extensive flake removal to shape the core in preparation for the main flake removals[32,51–55]. Part of the problem here is the use of the six criteria as a checklist rather than a guide. Boëda himself follows the checklist approach and defines cores as non-Levallois when one criterion is absent[56]. In more recent research, we see a move away from this checklist system and instead the adoption of a more holistic approach, using the criteria as a guide[41,57,58].

We follow this more holistic approach, identifying the Levallois in the Guanyindong Cave assemblage as large and flat preferential flakes, sometimes showing faceted platforms, and cores with hierarchical relationships and preferential removals. We do not require traces of extensive shaping, instead following previous work that recognizes naturally asymmetric surfaces as compatible with an identification of Levallois technology. The detailed analysis of Levallois elements in Guanyindong Cave lithic and the previously published results are summarized in the Supplementary Information.

OSL dating. OSL dating provides an estimate of the time since mineral grains such as quartz or feldspars were last exposed to sunlight[59–61]. The burial age is estimated by dividing the equivalent dose (D_e, a measure of the radiation energy absorbed by grains during their period of burial) by the environmental dose rate (the rate of supply of ionizing radiation to the grains over the burial period). Here we determined the sedimentary ages of our sediment samples based on the measurements of the OSL from quartz.

A total of 13 sediment samples were collected for OSL dating from two residual profiles (S1 and S2) at the south wall of the cave entrance (Extended Data Fig. 2), including three samples from layer 1 at S1, four from layer 2 at S2, two from layer 4 at S1 and one from each of the layers 5–8 at S1 (Extended Data Figs. 3, 4). We did not take any sample from Layer 3, because we could not find suitable materials for dating from S1 (see Extended Data Fig. 3). The samples were collected by hammering opaque plastic tubes, each about 5 cm in diameter and around 25 cm long, into the cleaned section face. The tubes were sealed in black plastic bags for safe transport. Apart from the tubes, additional sediment at each sample location was collected and placed in plastic zip-lock bags for measuring their current moisture contents and radioactivity.

The sample tubes were opened and prepared under dim red light in the OSL dating laboratory at the University of Wollongong. The materials at both ends of each tube were discarded because they might have been exposed to sunlight at the time of sample collection. Because insufficient feldspar grains were extracted from our samples, only quartz grains were measured. Quartz grains were extracted using standard preparation procedures[62]. First, the samples were dissolved in 10% hydrochloric acid to remove carbonate before they were subsequently treated with 30% hydrogen peroxide solution to remove organic matter. The remaining sample was dried and then sieved to isolate grains of 90–125, 90–150, 90–180 and 180–212 μm in diameter. Quartz grains were separated from other minerals by density separation using sodium polytungstate solutions of 2.62 and 2.75 specific gravities. The separated quartz grains were etched with 48% hydrofluoric acid for around 40 min to remove the alpha-irradiated rind of each quartz grain and to destroy any remaining feldspars. The etched grains were then rinsed in hydrochloric acid to remove any precipitated fluorides, before being dried and sieved again. All of the samples were dominated by silt (<63 μm), and a limited amount of 180–212 μm quartz grains were extracted from our samples. Therefore, apart from the limited number of 180–212-μm grains, we also determined the D_e using smaller grains (in the range of 90–180 μm) for each sample.

The environmental dose rate for etched quartz is mainly attributable to beta and gamma radiation, from the decay of ^{238}U, ^{235}U, ^{232}Th (and their daughter products) and ^{40}K in the deposits surrounding the dated grains and cosmic rays. Beta dose rates were measured directly by low-level beta counting of dried, homogenized and powdered sediment samples from the dosimetry bags, using a GM-25-5 multicounter system[63]. Gamma dose rates were measured at each sample location by an in situ gamma spectrometer, to account for any spatial heterogeneity in the gamma radiation field within 30 cm of each OSL sample. To accommodate the gamma detector, after removing the plastic sample tubes, we further drilled the holes to a depth of 30 cm using a hand auger. A two-inch (five-cm diameter) probe was inserted into the hole, and counts were collected for 60 min with a two-inch Na(Tl)

© 2019 Springer Nature Limited. All rights reserved.

crystal. The detector was calibrated using the concrete blocks at Oxford[64] and the gamma dose rate was determined using the 'threshold' technique[65]. The cosmic-ray dose rates were estimated according to a previously published method[66], based on the geomagnetic latitude and altitude of the Guanyindong site, as well as the thickness of sediment above each sample. Because our samples were collected from the cave entrance, we also allowed for the overhead limestone shielding and the configuration of the cave, by making a correction for the zenith angular distribution of cosmic rays[67]. We assigned a relative uncertainty of 10% to account for the systematic uncertainty in the primary cosmic-ray intensity. Because the cosmic ray constitutes only 1–5% of the total dose rate for these samples (Supplementary Table 5), the OSL ages are not highly sensitive to errors associated with the cosmic-ray dose rate.

Each of the measured beta and gamma dose rates and the calculated cosmic-ray dose rate were corrected for attenuation by water. For the samples from S1, the measured water contents of the six samples from group B range from 20% to 24% (with a mean value of 22%) (Supplementary Table 5), but lower values (11–17%) were obtained for the three samples from layer 1. By contrast, higher values (28–32%) were found for all the samples taken from group A at S2. The difference in the water contents between the two profiles is expected as S1 has been exposed for several decades after the last excavation in 1970s, so the measured present-day water contents should be underestimated. By contrast, S2 was protected by stones and covered by vegetation, which should retain water content better than S1. We, therefore, expect that the water content obtained from S2 is more representative to the long-term water content of S1. To assess the water content more reliably, we took additional sedimentary samples from two of the original trenches (profile 2a and 3) inside the cave, where moisture contents are also better retained. For the 15 samples (with burial depth ranging from around 50 to approximately 300 cm) that we measured, water contents ranged from 15 to 40%, and the mean and standard deviation were 30% and 8.5%, respectively. So, instead of using the in situ water content, we used a value of 30% as an estimate of the long-term water content for our OSL samples from groups A and B and a value of 20% for those from layer 1. We assigned a 25% relative standard error to these estimates, to accommodate any likely variations in the water content over the burial period. We noted that the measured in situ water contents are within the 2σ range of the assumed values.

OSL measurements were made on an automated Risø TL-DA-20 luminescence reader equipped with a single-grain laser (532 nm)[68]. Laboratory irradiations were carried out within the luminescence reader using a calibrated $^{90}Sr/^{90}Y$ beta source. All the quartz OSL measurements were made by mounting the grains onto standard Risø single-grain discs (gold-plated aluminium discs drilled with 100 holes that are each 300 mm in diameter and 300 mm deep)[69], where each grain hole contained one grain of 180–212 μm in diameter, or about eight grains of 90–125 μm in diameter. Spatial variation in the dose rate for individual grain positions was calibrated using gamma-irradiated quartz standards from the instrument manufacturer Risø. The ultraviolet OSL emissions were detected by an Electron Tubes 9235QA photomultiplier tube fitted with Hoya U-340 filters.

All OSL measurements were made using a single-aliquot regenerative-dose (SAR) procedure[70,71]. The SAR procedure involves measuring the OSL signals from the natural (burial) dose and from a series of regenerative doses, each of which was preheated at 240 °C for 10 s before optical stimulation by the green laser beam for 2 s at 125 °C. A fixed test dose (around 16 Gy) was given after each natural and regenerative dose, and the induced test-dose OSL signals were used to correct for any sensitivity changes during the SAR sequence. A cut heat to 180 °C was applied to the test dose. A duplicate regenerative dose was included in the procedure, to check on the validity of sensitivity correction, and a 'zero dose' measurement was made to monitor the extent of any 'recuperation' or 'thermal transfer' induced by the 240 °C preheating step. As a check on possible contamination from feldspars, we also applied the OSL infrared depletion-ratio test[72] at the end of the SAR sequence, using an infrared bleach of 40 s at 50 °C.

To test whether the SAR procedure is suitable for our samples, a dose-recovery test was conducted on sample GYD-OSL2 using different combinations of preheat/cutheat (260/180, 240/180, 220/180, 200/160 and 180/160 °C) temperatures. Two single-grain discs were measured for each preheat temperature using the grains of 90–125 μm diameter. The grains were bleached for approximately 30 min using a Dr Hönle solar simulator (model: UVACUBE 400). The bleached grains were then given a dose of around 100 Gy, before being measured using the SAR procedure using different preheat and cutheat temperatures. To select reliable single-grain D_e results, we applied several rejection criteria similar to those proposed previously[73]. Grains were rejected if they exhibited one or more of the following properties. (1) Test-dose signal (T_n) too dim, that is, the initial intensity is below the instrument detection limit (3σ below the background intensity) and/or the relative standard error on the test-dose measurement was more than 20%. (2) High levels of recuperation (that is, the ratio between the sensitivity-corrected OSL signals for the zero dose and the largest regenerative dose is higher than 5%). (3) Poor dose–response curve (DRC), that is, the regenerative signals are too scattered to be well-fitted with suitable functions (for example, a linear or saturating exponential function); note that poor recycling ratio falls into this category. (4) Natural OSL signal statistically equal to or greater than the saturation level of the corresponding DRC.

For each of the preheat temperatures, 39–64 grains were accepted after applying the above rejection criteria. The measured to given dose ratios (or dose-recovery ratios) are summarized as radial plots in Supplementary Fig. 1a–e for each of the preheat temperatures. We applied a central age model[70] to calculate the weighted mean recovery ratios for each preheat temperature, and these are shown in each of the radial plots[70,74]. The dose-recovery results are plotted against the preheat temperature in Supplementary Fig. 1f. The mean ratios are statistically consistent with unity at 1σ for the preheat temperatures at 220, 240 and 260 °C, which suggests that the chosen SAR procedures can accurately recover a known dose under these conditions.

On the basis of the dose-recovery tests, we chose the preheat/cutheat of 240/180 °C for measuring D_e values for all samples. Supplementary Fig. 1g, h shows the natural OSL decay curves of 10 grains each for GYD-OSL2 and GYD-OSL6. On the basis of the measurements from the 180–212-μm diameter grains, we found that the OSL intensity varies significantly from grain to grain, and most (around 90%) of the grains yielded no OSL signal at all (or their signal intensity was below the instrumental limit of detection); fewer than 5% of the measured single grains contributes >90% of the total OSL signal (Supplementary Fig. 1i). Apart from the OSL intensity, the DRCs from different grains also display a wide range of shapes associated with different saturation doses (Supplementary Figs. 1–20).

Depending on the availability of separated grains, 800–4,200 grains of 180–212 μm diameter were measured for GYD-OSL1, 2, 3, 5 and 6 (Supplementary Table 3). However, only about 2% of measured grains could pass the rejection criteria described above, and about 90% of the grains were rejected because the signals were too weak. For this reason, we measured smaller grains in the range of 90–180 μm for all of the samples. For the measurement of small grain size (<180 μm diameter) fractions, each grain hole of the standard single-grain disc may contain several grains (for example, up to eight grains of 90–125 μm diameter), which makes our measurements equivalent to a small aliquot that contains only a few grains. There are several advantages of measuring smaller grains. First, several grains were measured together in each of the holes, so there is a higher probability to find a bright grain in each hole, providing a considerable reduction in instrument time. Second, because of the low percentage (<5%) of bright grains in our samples, the measured OSL signal from each of the grain holes is expected to be dominated by only one or two grains, thereby effectively making these measurements equivalent to single-grain measurements. This is further confirmed by the similar results obtained from the 180–212-μm diameter grains and smaller grains (Supplementary Table 5). Using this method, 500–1,400 small aliquots were measured for each of the samples (Supplementary Table 3). As expected, the percentage of aliquots that have detectable OSL signals was significantly increased, ranging from 18% to 55%. About 20% of the small aliquots produced more than 80% of the total OSL signal (Supplementary Fig. 1i). Correspondingly, the proportion of grains that passed the rejection criteria was considerably increased (Supplementary Table 3).

The distributions of individual D_e values that passed the rejection criteria are shown in radial plots in Supplementary Fig. 2 for all of the samples. All of the samples show a large range in D_e values, ranging from around 0 to about 250 Gy. For those samples for which two grain sizes were measured, similar D_e distributions were observed between the two grain sizes from the same sample. These broad D_e distributions indicate that our samples were contaminated by 'younger' grains, especially in the case of the samples taken from S1. This is not surprising, because the residual profiles have been exposed for several decades since the last excavation in 1970s. As a result, one would expect some degree of bioturbation that could have intruded younger grains into the profiles. Evidence of such post-depositional mixture can be seen from the modern tree roots that penetrate deeply into the profile as shown in Extended Data Fig. 3. Fortunately, such recent bioactivity did not destroy the stratigraphic integrity of the residual profiles, because clear sedimentary beddings are still visible (Extended Data Figs. 3, 4) and these are consistent with the description in the original excavation report.

The numbers of grains or aliquots that were rejected based on each of the rejection criteria are summarized in Supplementary Table 3. There are considerable proportions of grains or aliquots (up to around 40%) that have saturated natural signals, for example, the L_n/T_n value is statistically consistent to or above the saturation level of the corresponding DRCs. As a result, finite D_e estimates cannot be obtained for these grains. Recent studies have suggested that rejecting a large number of 'saturated' grains may result in a significant underestimation of the final D_e estimate due to the truncation of the full D_e distribution[75–79]. To avoid this problem, a new method[80] has been proposed for the analysis of the L_n/T_n distribution and to establish standardized growth curves (SGCs)[81,82] for different grains or aliquots. Using this new method, no grains were rejected because they were 'saturated' and, therefore, a full and untruncated distribution of the L_n/T_n ratios was obtained, which enables reliable D_e estimation beyond the conventional

© 2019 Springer Nature Limited. All rights reserved.

limit of approximately $2D_0$ using the standard SAR procedure. Given the large proportion of 'saturated' grains in our samples, we therefore, applied this method[80] to estimate D_e values for our samples.

We first investigated the variability in the DRCs for our samples and the possibility of establishing SGCs following the previously proposed method[79]. By analysing the L_x/T_x ratios between two regenerative doses, it was previously[79] found that the single-grain and small-aliquot DRCs could be divided into three broad groups termed 'early', 'medium' and 'later', which saturated at different dose levels. It was also shown that each group could be well-defined by a SGC. As suggested previously[79], SGCs should be established using only those aliquots (grains) that are considered to be well-behaved so that reliable growth curves are produced. To do this, we first identified and rejected poorly behaved grains or aliquots using similar rejection criteria to those mentioned above but included all the 'saturated' ones. Supplementary Figure 3a shows comparisons of all the DRCs that pass the rejection criteria for the 90–150-μm quartz grains from GYD-OSL1. The DRCs from the same samples are highly variable among different grains or aliquots, which prevents the establishment of a common SGC. To test whether the samples can be classified into several groups that share the same DRCs, we calculated the ratios between the L_x/T_x values of two regenerative doses of around 280 and about 70 Gy, which reflects the saturation dose levels of the corresponding DRCs[79], for example, higher ratios represent larger saturation doses or later saturation. The ratios are shown in the radial plots in Supplementary Fig. 3b.

To test whether there are several groups that each have a similar saturation dose, we used a finite mixture model (FMM)[74,83,84] to identify the number of groups that have statistically indistinguishable L_x/T_x ratios and estimate the weighted mean ratios for each group and the probability of falling in each group for each grain or aliquot (Supplementary Fig. 3b). The DRCs from each group were analysed using a least-square normalization procedure[79] to establish corresponding SGCs for each of the groups (Supplementary Fig. 3c). The dose–response data from the same groups were fitted using a general-order kinetic function[85] of the form $f(x) = a(1-(1+bcx)^{(-1/c)}) + d$, where x is the dose and parameters a, b, c and d are constants. The different groups have considerably different saturation dose levels, that is, group 1 saturated at around 100 Gy, whereas group 3 showed no sign of saturation up to 500 Gy. The ratio between the measured L_x/T_x and the expected values based on the SGC are statistically consistent with unity for all of the groups; most of these ratios (around 90% or more) are consistent with unity at 2σ (Supplementary Fig. 3d–f), confirming the validity of the grouping and SGC establishment. The same procedure was applied to all of our samples, and we found that most of our samples could be fitted to 2–4 groups (Supplementary Figs. 3–20) despite the large variation in DRCs observed.

Once the SGCs were established for individual groups, the natural signals (L_n/T_n) from each of the groups were renormalized using the same scaling factors obtained during the least-square normalization procedure. The distributions of the least-square-normalized L_n/T_n values for each of the groups that were used to calculate final D_e values for each sample are shown in Supplementary Figs. 3–20 for all the samples. All groups were dominated by a single population, although most of them contain a few grains that have significantly smaller L_n/T_n values. This is similar to the patterns observed in the distribution of the SAR D_e values (Supplementary Fig. 2). However, because all of the grains that were rejected due to 'saturation' are included, it appears that all samples have a dominant population and this population has the highest L_n/T_n (or D_e) values. Therefore, we consider that the dominant population represents the true natural doses of the grains that remained intact since their burial.

The single-grain DRCs, SGCs and distribution of L_n/T_n values for individual groups of different samples are shown in Supplementary Figs. 3–20. For samples showing a single population of L_n/T_n values, we applied a central age model to estimate the weighted mean L_n/T_n values. For those with only a few young grains introduced, we identified and removed these outliers based on the median absolute deviation as a means of screening data for outliers[86,87]. For these cases, we calculated the normalized median absolute deviation using 1.4826 as the appropriate correction factor for a normal distribution, and rejected $\log(L_n/T_n)$ values with a normalized median absolute deviation greater than 1.5. For the other samples for which discrete D_e components could clearly be identified and are statistically supported, we applied the FMM to identify the number of populations for each distribution of least-square-normalized L_n/T_n and to calculate the central value of each population. The FMM was fitted by varying the common overdispersion value (σ_b) between 0 and 0.5 to find the optimum fit when the lowest Bayes Information score was reached[74,88]. The best-fit overdispersion values (or σ_b) for FMM fell within 0.1–0.2 for all samples. The best estimates of the least-square-normalized L_n/T_n for each group were then projected onto the corresponding SGCs to estimate their D_e. The D_e results for all of the samples are summarized in Supplementary Table 4. For some samples (for example, the 180–212 μm grains of sample GYD-OSL5), insufficient number of grains were accepted, so reliable results cannot be obtained. Group 1 (that is, the early saturated group) of most samples yielded infinite D_e values, because the L_n/T_n values were statistically within the saturation level of the corresponding SGC. However, finite results were obtained for the other groups that had higher saturation doses and their D_e values are statistically indistinguishable from each other for the same sample. For the samples for which two different grain sizes were measured, the D_e values from the two fractions were statistically consistent with each other. These results further confirm the validity of the grouping, SGC establishment and D_e estimates based on L_n/T_n and SGCs.

We estimated the D_e values for each grain size fraction of the samples based on the weighted mean of the results for the non-saturated DRC groups that produced finite D_e values. The final D_e and age estimates for the GYD samples are listed in Supplementary Table 5, together with the dose-rate estimates. For the samples for which two grain sizes were measured, the ages obtained from both grain sizes are consistent with each other within 1σ, further supporting our argument that the small-aliquot measurements are analogue to single-grain measurements. Therefore, for the samples for which two different grain sizes were measured, we estimated their ages based on the weighted mean of the ages obtained from the two grain sizes. The final age estimates for all the samples are shown in Fig. 2 and Supplementary Table 5.

Our OSL chronology provides a firm constraint on the sedimentary ages of the artefact-bearing deposits from layer 1, groups A and B. The OSL age for the sample from layer 6 is consistent with the U-series age (around 180 kyr) of the stalagmite sample taken from the same layer (Supplementary Table 1), confirming the reliability of both dates. On the basis of the new OSL ages and previous U-series dating results (see Supplementary Information for a full discussion on U-series results), we conclude that layer 2 (group A) was deposited around 80–90 ka, corresponding to the last interglacial period or MIS 5a. Our age estimate for group A is further supported by sedimentary features. The deposits of group A consist of reddish clay and are indicative of strong paedogenesis process taking place during warm and humid interglacial conditions. The poorly preserved fossils in group A, compared to those in group B, further support that the depositional environment of group A was relatively warm and humid. Layers 4–8 (group B) were deposited between 160 and 170 ka. The age of the Guanyindong lithic assemblage can, therefore, be safely constrained to between approximately 170 and 80 ka.

Code availability. All custom R scripts used to produce the results presented here are available online at https://doi.org/10.17605/OSF.IO/ERNTJ.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data are available from the corresponding authors upon reasonable request.

31. Boëda, E. in *The Definition and Interpretation of Levallois Technology* (eds Dibble, H. & Bar-Yosef, O.) 41–68 (Prehistory Press, Madison, 1995).
32. Brantingham, P. J. & Kuhn, S. L. Constraints on Levallois core technology: a mathematical model. *J. Archaeol. Sci.* **28**, 747–761 (2001).
33. Eren, M. I. & Lycett, S. J. Why Levallois? A morphometric comparison of experimental 'preferential' Levallois flakes versus debitage flakes. *PLoS ONE* **7**, e29273 (2012).
34. White, M., Ashton, N. & Scott, B. in *The Ancient Human Occupation of Britain* (eds Ashton, N. et al.) 53–66 (Elsevier, Amsterdam, 2011).
35. Van Peer, P. *The Levallois reduction strategy* (Prehistory Press, Madison, 1992).
36. Schlanger, N. Understanding Levallois: lithic technology and cognitive archaeology. *Camb. Archaeol. J.* **6**, 231–254 (1996).
37. Mellars, P. A. *The Neanderthal Legacy: An Archaeological Perspective from Western Europe* (Princeton Univ. Press, Princeton, 1995).
38. Monnier, G. F. & Missal, K. Another Mousterian Debate? Bordian facies, chaîne opératoire technocomplexes, and patterns of lithic variability in the western European Middle and Upper Pleistocene. *Quat. Int.* **350**, 59–83 (2014).
39. Malinsky-Buller, A. The muddle in the Middle Pleistocene: the Lower–Middle Paleolithic transition from the Levantine perspective. *J. World Prehist.* **29**, 1–78 (2016).
40. Barzilai, O., Malinsky-Buller, A. & Ackermann, O. Kefar Menachem West: a Lower Paleolithic site in the southern Shephela, Israel. *J. Israel Prehist. Soc.* **36**, 7–38 (2006).
41. White, M. & Ashton, N. Lower Palaeolithic core technology and the origins of the Levallois method in north-western Europe. *Curr. Anthropol.* **44**, 598–609 (2003).
42. Niu, D. et al. The initial Upper Palaeolithic in northwest China: new evidence of cultural variability and change from Shuidonggou locality 7. *Quat. Int.* **400**, 111–119 (2016).
43. Shimelmitz, R., Weinstein-Evron, M., Ronen, A. & Kuhn, S. L. The Lower to Middle Paleolithic transition and the diversification of Levallois technology in the Southern Levant: evidence from Tabun Cave, Israel. *Quat. Int.* **409**, 23–40 (2016).
44. Brantingham, P. J., Kuhn, S. L. & Kerry, K. W. *The Early Upper Paleolithic beyond Western Europe* (California Univ. Press, 2004).
45. Cahen, D. Les industries préhistoriques des nappes alluviales de Petit-Spiennes et de Mesvin. *Notae Praehistoricae* **1**, 70–74 (1981).
46. Cahen, D., Haesaerts, P. & Watteyne, D. La nappe alluviale de Petit-Spiennes et le début du débitage levallois dans la vallée de la Haine. *Archaeologia Belgica Bruxelles* **1**, 7–16 (1985).

© 2019 Springer Nature Limited. All rights reserved.

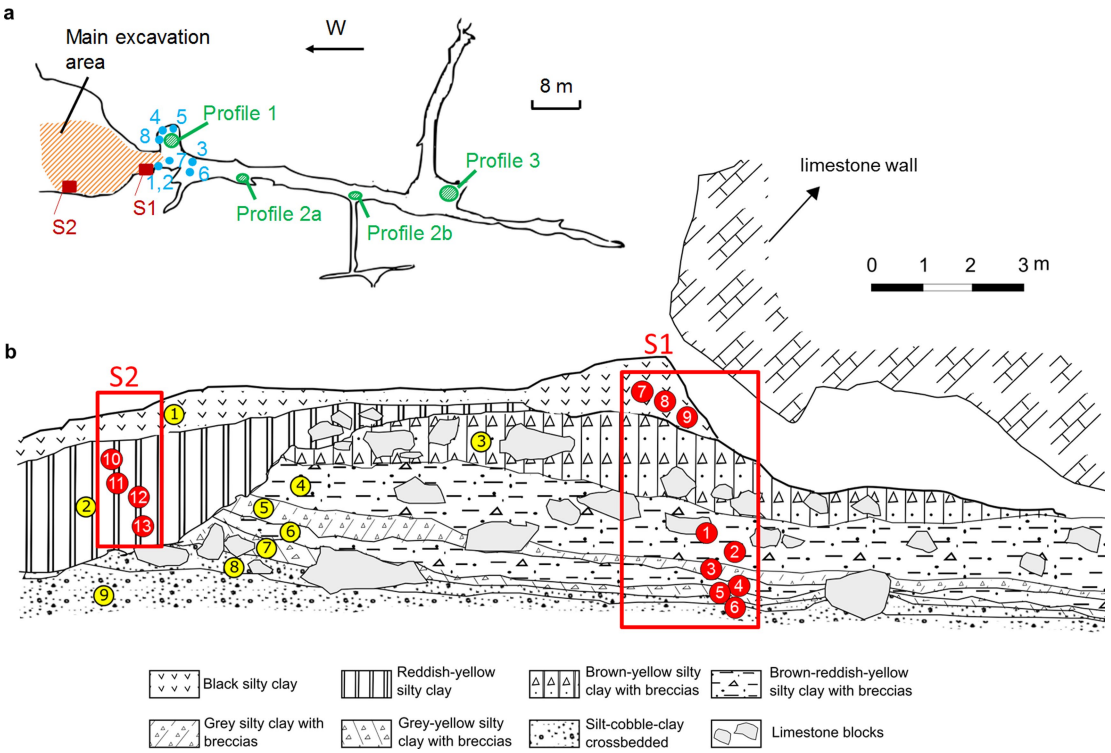
47. Watteyne, D. Petit-Spiennes: industrie (s) a débitage Levallois et para-Levallois. *Notae Praehistoricae* **5**, 95–104 (1985).
48. Roe, D. A. *The Lower and Middle Palaeolithic Periods in Britain Vol*. **46** (Routledge, London, 2014).
49. Otte, M. in *The Definition and Interpretation of Levallois technology* (eds Dibble, H. & Bar-Yosef, O.) 117–124 (Prehistory Press, Madison, 1995).
50. Ryssaert, C. Some new insights in an old collection: lithic technology at Mesvin IV. *Notae Praehistoricae* **26**, 91–99 (2006).
51. Kuhn, S. L. *Mousterian Lithic Technology* (Princeton Univ. Press, Princeton, 1995).
52. Delagnes, A. in *The Definition and Interpretation of Levallois technology Monographs in World Archaeology* Vol. 23 (eds Bar-Yosef, O. & Dibble, H.) 201–212 (Prehistory Press, Madison, 1995).
53. Chazan, M. Redefining Levallois. *J. Hum. Evol*. **33**, 719–735 (1997).
54. Picin, A. Technological adaptation and the emergence of Levallois in Central Europe: new insight from the Markkleeberg and Zwochau open-air sites in Germany. *J. Quat. Sci*. **33**, 300–312 (2018).
55. Boëda, E. & Pelegrin, J. Approche technologique du nucleus levallois a éclat. *Etudes Préhistoriques Lyon* **15**, 41–48 (1979).
56. Boëda, E. Approche technologique du concept Levallois et évaluation de son champ d'application: étude de trois gisement saaliens et weichseliens de la France septentrionale. PhD thesis, Université de Paris X (1986).
57. Scott, R. The Early Middle Palaeolithic of Britain; origins, technology and landscape. PhD thesis, Durham University (2006).
58. Bolton, L. Assessing the Origins of Levallois through Lower Palaeolithic Core Variation: A Comparative Study of Simple Prepared Cores in Northwest Europe. PhD thesis, University of Southampton (2015).
59. Huntley, D. J., Godfrey-Smith, D. I. & Thewalt, M. L. W. Optical dating of sediments. *Nature* **313**, 105–107 (1985).
60. Aitken, M. J. *An Introduction to Optical Dating* (Oxford Univ. Press, Oxford, 1998).
61. Roberts, R. G. et al. Optical dating in archaeology: thirty years in retrospect and grand challenges for the future. *J. Archaeol. Sci*. **56**, 41–60 (2015).
62. Wintle, A. G. Luminescence dating: laboratory procedures and protocols. *Radiat. Meas*. **27**, 769–817 (1997).
63. Bøtter-Jensen, L. & Mejdahl, V. Assessment of beta dose-rate using a GM multicounter system. *Int. J. Rad. Appl. Instrum. D* **14**, 187–191 (1988).
64. Rhodes, E. J. & Schwenninger, J.-L. Dose rates and radioisotope concentrations in the concrete calibration blocks at Oxford. *Ancient TL* **25**, 5–8 (2007).
65. Mercier, N. & Falguères, C. Field gamma dose-rate measurement with a NaI(Tl) detector: re-evaluation of the "threshold" technique. *Anc. TL* **25**, 1–4 (2007).
66. Prescott, J. R. & Hutton, J. T. Cosmic-ray contributions to dose rates for luminescence and ESR dating: large depths and long-term time variations. *Radiat. Meas*. **23**, 497–500 (1994).
67. Smith, M. A., Prescott, J. R. & Head, M. J. Comparison of ^{14}C and luminescence chronologies at Puritjarra rock shelter, central Australia. *Quat. Sci. Rev*. **16**, 299–320 (1997).
68. Bøtter-Jensen, L., Andersen, C. E., Duller, G. A. T. & Murray, A. S. Developments in radiation, stimulation and observation facilities in luminescence measurements. *Radiat. Meas*. **37**, 535–541 (2003).
69. Bøtter-Jensen, L., Bulur, E., Duller, G. A. T. & Murray, A. S. Advances in luminescence instrument systems. *Radiat. Meas*. **32**, 523–528 (2000).
70. Galbraith, R. F., Roberts, R. G., Laslett, G. M., Yoshida, H. & Olley, J. M. Optical dating of single and multiple grains of quartz from Jinmium rock shelter, northern Australia: part 1, experimental design and statistical models. *Archaeometry* **41**, 339–364 (1999).
71. Murray, A. S. & Wintle, A. G. Luminescence dating of quartz using an improved single-aliquot regenerative-dose protocol. *Radiat. Meas*. **32**, 57–73 (2000).
72. Duller, G. A. T. Distinguishing quartz and feldspar in single grain luminescence measurements. *Radiat. Meas*. **37**, 161–165 (2003).
73. Jacobs, Z., Duller, G. A. T. & Wintle, A. G. Interpretation of single grain D_e distributions and calculation of D_e. *Radiat. Meas*. **41**, 264–277 (2006).
74. Galbraith, R. F. & Roberts, R. G. Statistical aspects of equivalent dose and error calculation and display in OSL dating: an overview and some recommendations. *Quat. Geochronol*. **11**, 1–27 (2012).
75. Gliganic, L. A., Jacobs, Z., Roberts, R. G., Domínguez-Rodrigo, M. & Mabulla, A. Z. P. New ages for Middle and Later Stone Age deposits at Mumba rockshelter, Tanzania: optically stimulated luminescence dating of quartz and feldspar grains. *J. Hum. Evol*. **62**, 533–547 (2012).
76. Duller, G. A. T. Improving the accuracy and precision of equivalent doses determined using the optically stimulated luminescence signal from single grains of quartz. *Radiat. Meas*. **47**, 770–777 (2012).
77. Thomsen, K. J. et al. Testing single-grain quartz OSL methods using sediment samples with independent age control from the Bordes-Fitte rockshelter (Roches d'Abilly site, Central France). *Quat. Geochronol*. **31**, 77–96 (2016).
78. Guo, Y.-J. et al. New ages for the Upper Palaeolithic site of Xibaimaying in the Nihewan Basin, northern China: implications for small-tool and microblade industries in north-east Asia during Marine Isotope Stages 2 and 3. *J. Quat. Sci*. **32**, 540–552 (2017).
79. Li, B., Jacobs, Z. & Roberts, R. G. Investigation of the applicability of standardised growth curves for OSL dating of quartz from Haua Fteah cave, Libya. *Quat. Geochronol*. **35**, 1–15 (2016).
80. Li, B., Jacobs, Z., Roberts, R. G., Galbraith, R. & Peng, J. Variability in quartz OSL signals caused by measurement uncertainties: problems and solutions. *Quat. Geochronol*. **41**, 11–25 (2017).
81. Roberts, H. M. & Duller, G. A. T. Standardised growth curves for optical dating of sediment using multiple-grain aliquots. *Radiat. Meas*. **38**, 241–252 (2004).
82. Li, B., Roberts, R. G., Jacobs, Z. & Li, S. H. Potential of establishing a 'global standardised growth curve' (gSGC) for optical dating of quartz from sediments. *Quat. Geochronol*. **27**, 94–104 (2015).
83. Roberts, R. G., Galbraith, R. F., Yoshida, H., Laslett, G. M. & Olley, J. M. Distinguishing dose populations in sediment mixtures: a test of single-grain optical dating procedures using mixtures of laboratory-dosed quartz. *Radiat. Meas*. **32**, 459–465 (2000).
84. Galbraith, R. F. & Green, P. F. Estimating the component ages in a finite mixture. *Int. J. Rad. Appl. Instrum. D* **17**, 197–206 (1990).
85. Guralnik, B. et al. Radiation-induced growth and isothermal decay of infrared-stimulated luminescence from feldspar. *Radiat. Meas*. **81**, 224–231 (2015).
86. Rousseeuw, P. J. & Croux, C. Alternatives to the median absolute deviation. *J. Am. Stat. Assoc*. **88**, 1273–1283 (1993).
87. Rousseeuw, P. J., Debruyne, M., Engelen, S. & Hubert, M. Robustness and outlier detection in chemometrics. *Crit. Rev. Anal. Chem*. **36**, 221–242 (2006).
88. Arnold, L. J. & Roberts, R. G. Stochastic modelling of multi-grain equivalent dose (D_e) distributions: implications for OSL dating of sediment mixtures. *Quat. Geochronol*. **4**, 204–230 (2009).

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 1 | Photos showing the landscape and location of the Guanyindong Cave. a, Southward view of the Guanyindong Cave. **b**, The main entrance of the cave.

© 2019 Springer Nature Limited. All rights reserved.



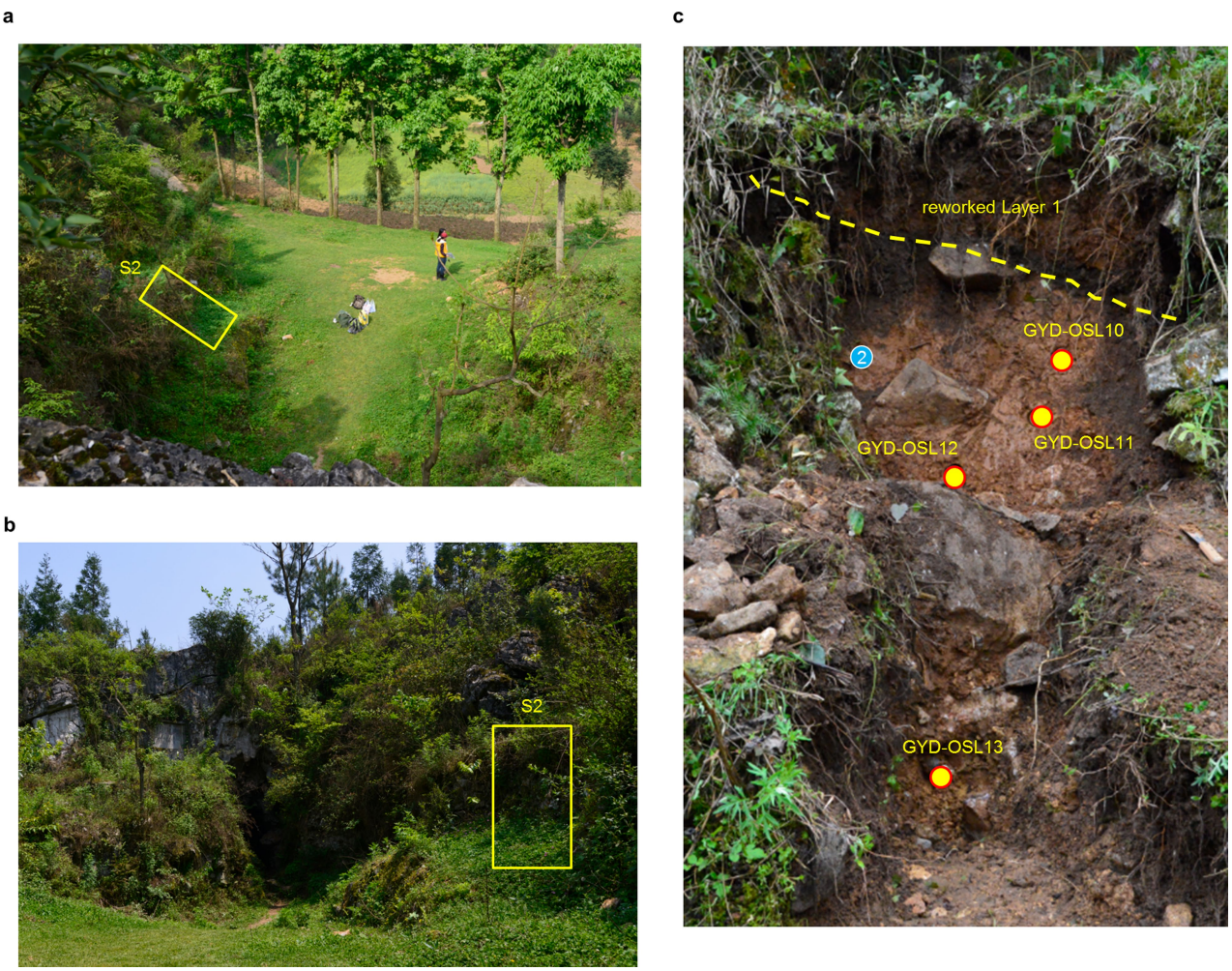
Extended Data Fig. 2 | Plan view and stratigraphy of the Guanyindong Cave. a, Plan view of the cave, main excavation area and the residual profiles from the south wall. The blue dots and the numbers next to each of the dots represent the locations of U-series dating samples have been taken previously[17] (see Supplementary Information for discussion of the U-series results); sample codes from 1 to 8 are QGC-19-1, QGC-19-2, QGC-4, QGC-21, QGB-4, QGC-7 and QGC-23, respectively. The green circles are the locations of profiles 1, 2a, 2b and 3. The red squares show the locations of the residual profiles S1 and S2, where the OSL samples were taken. **b**, Detail of the numbered stratigraphic layers at the main entrance of the cave. The stratigraphic layer numbers are shown in yellow circles. The red rectangles show the locations of the two south-wall sections (S1 and S2) where OSL samples were taken. The locations of OSL samples are shown in red circles, with the sample code shown inside (for example, number 1 represents GYD-OSL1; see Extended Data Figs. 3, 4 for more details). **a**, **b**, Images were adapted from a previous study[16], copyright 1986.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | General view of the residual profile S1 from the cave entrance. a, Photo taken from the interior of the cave, showing the location of the residual profile S1 at the south wall (marked by a rectangle with details shown in **b** and **c**). **b**, Photo showing details of the residual profile S1 at the south wall and the location of all OSL samples from layer 1 and layers 4–8. The details of layers 3–9 inside the yellow rectangle are shown in **c**. **c**, Photo showing the details of sedimentary layers 3–9 of group B, and the location of OSL samples. The stratigraphic layer numbers are shown in blue circles and the location of OSL samples are marked by yellow circles with sample names shown next to each of them. The dashed yellow lines in **b** and **c** show the boundaries between the layers.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 4 | General view of the residual profile S2 outside the cave entrance. a, Photo taken from top of the cave, showing the location of the residual profile S2 (indicated by the rectangle). **b**, Photo taken from outside the cave, showing the location of the residual profile S2 (indicated by the rectangle). **c**, Photo showing the details of sedimentary layers (layer 2 and reworked layer 1) of residual profile S2, and the location of OSL samples. The dashed yellow line shows the boundary between layers 1 and 2. The stratigraphic layer numbers are shown in blue circles and the location of OSL samples are marked by yellow circles with sample names shown next to each of them.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Photographs of selected Levallois cores. a, **d**, **f**, Levallois recurrent cores. **b**, **c**, **e**, Levallois preferential cores. The line drawings of these artefacts are shown in Fig. 3a–f. The artefacts shown in **b** and **c** were recovered from group A.

© 2019 Springer Nature Limited. All rights reserved.



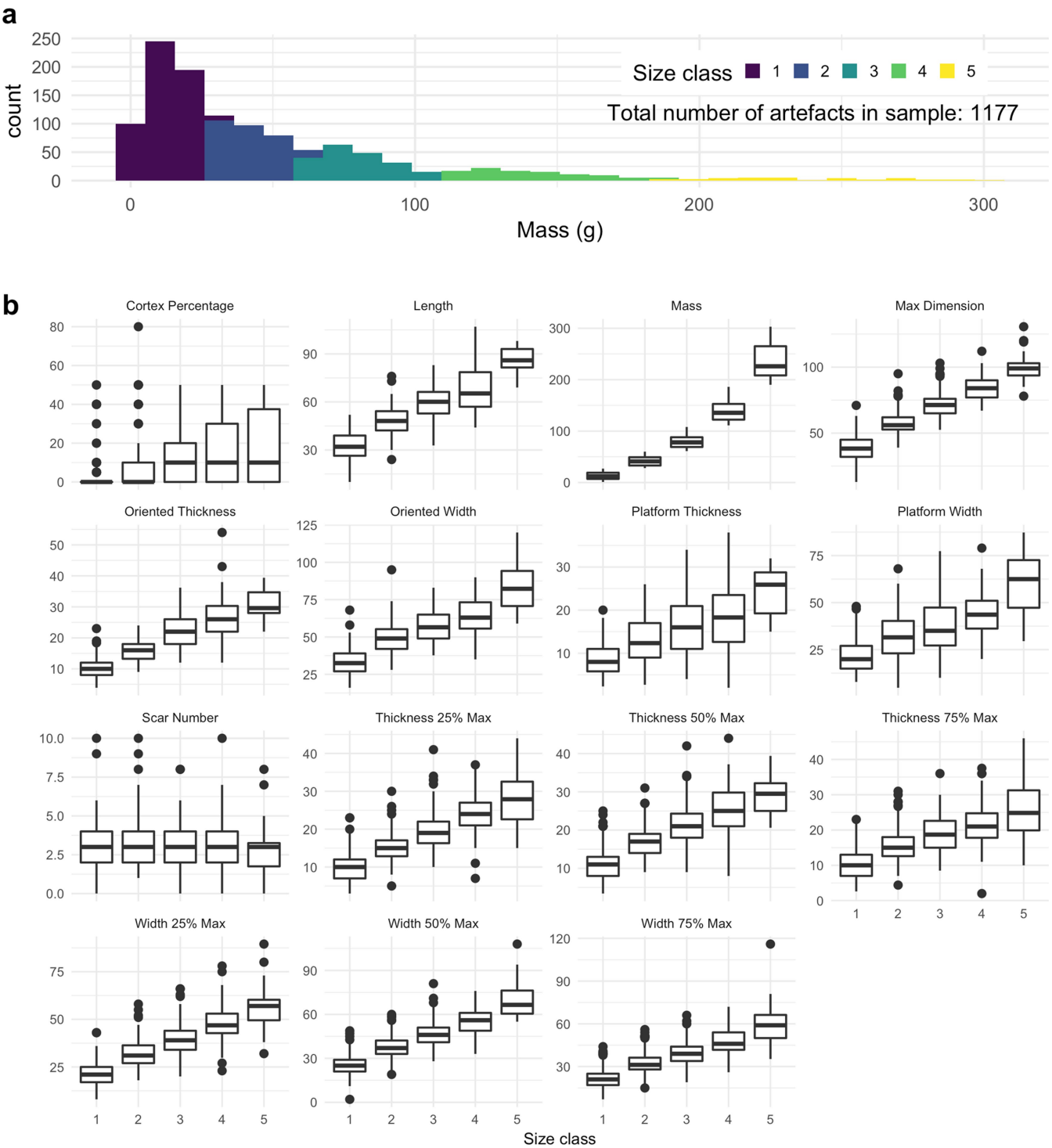
Extended Data Fig. 6 | Photographs of selected Levallois flakes and tools. g–k, **n**, Levallois flakes. **l**, Débordant. **m**, Tools made on Levallois blanks. **o**, **p**, Pseudo-Levallois points. The line drawings of these artefacts are shown in Fig. 3g–p.

© 2019 Springer Nature Limited. All rights reserved.



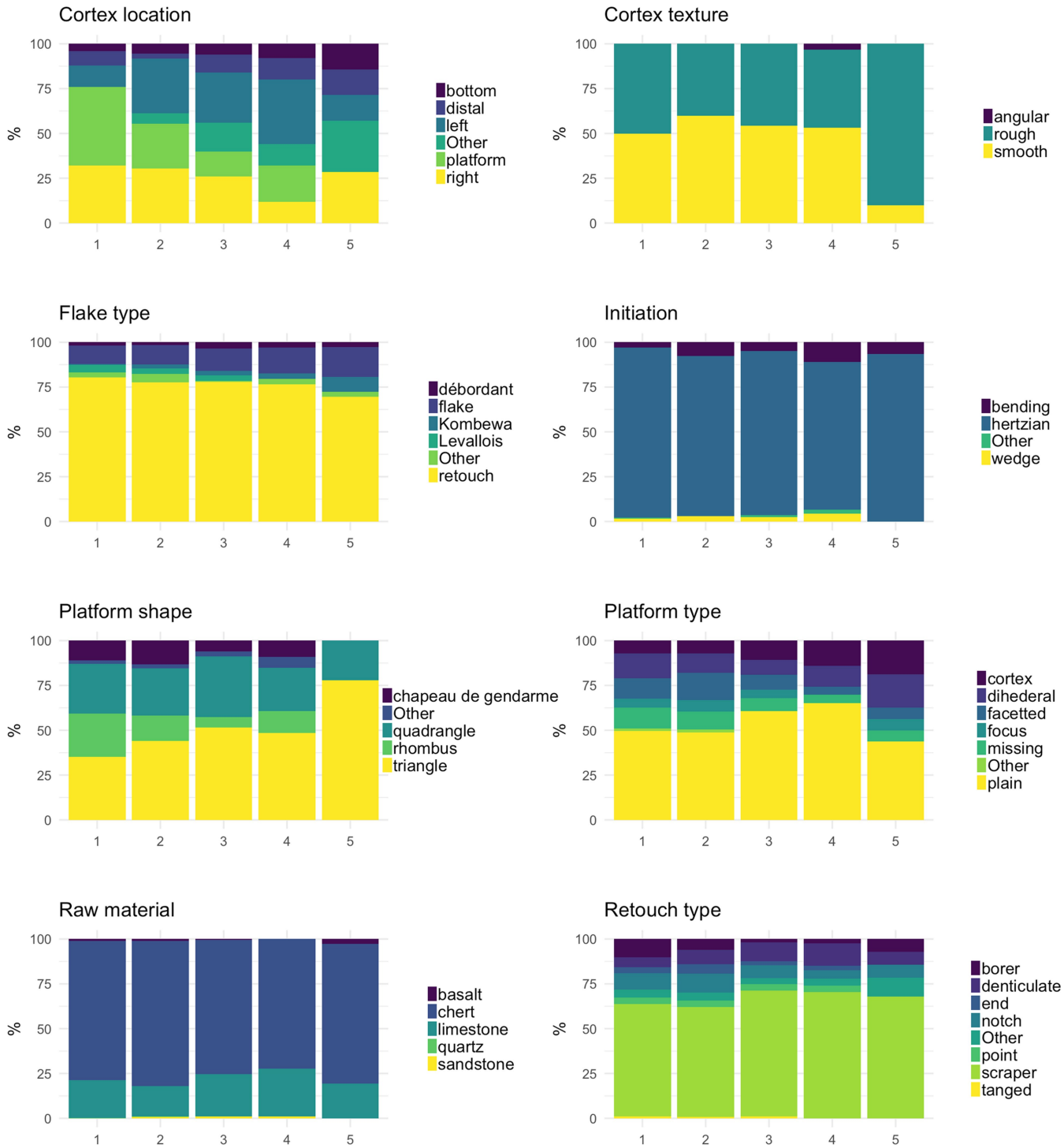
Extended Data Fig. 7 | Photographs of selected Levallois tools and flakes with prepared platform. q–s, Tools made on Levallois blanks. **t–z**, Flakes with prepared platforms. The line drawings of these artefacts are shown in Fig. 3q–z. The artefact shown in **q** was recovered from group A, and those shown in **r** and **s** were from group B.

© 2019 Springer Nature Limited. All rights reserved.



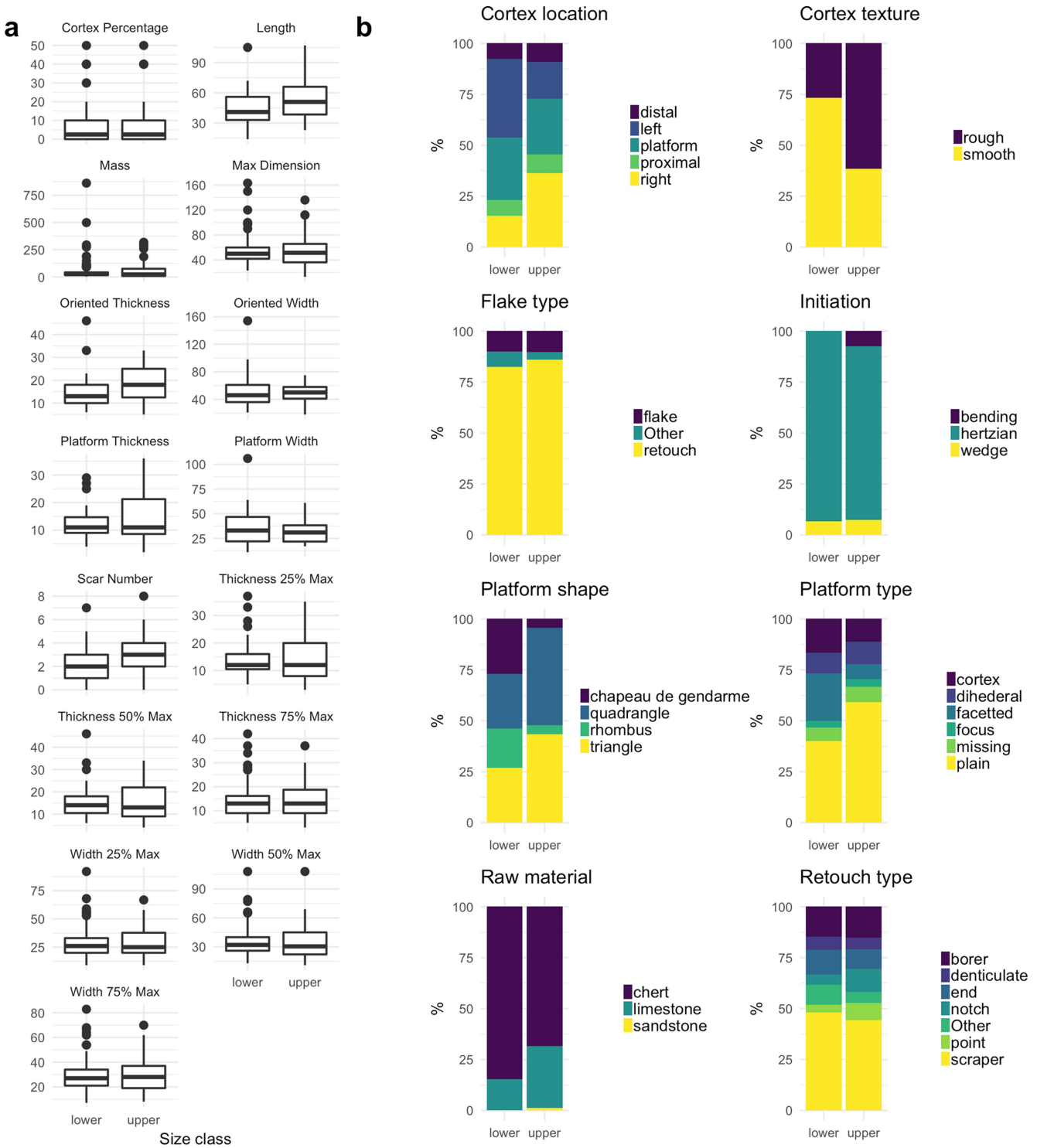
Extended Data Fig. 8 | Distributions of metric variables on flakes. **a**, Histogram of flake lengths, coloured by size class. **b**, Box-and-whisker plots of a selection of metric variables to show technological variation across the size classes to reveal the lithic reduction sequence ($n = 1{,}177$ flakes). Centre lines show data median, boxes show first and third quartiles (the 25th and 75th percentiles), and the whiskers extend from the upper and lower hinge to the largest and smallest values that are no further than 1.5 times the interquartile range from the hinge (which is the distance between the first and third quartiles). Data beyond the end of the whiskers are outlying points and are plotted individually. Linear dimensions are measured in mm, mass in g.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Distributions of technological attributes of flakes across the five size classes. $n = 1{,}177$ flakes.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Comparison of flakes from the upper (group A) and lower (group B) layers of the deposit ($n = 204$), with 117 pieces from the lower layers (dated to 170–160 ka) and 87 from the upper layer (dated to approximately 90–80 ka). a, Metric variables. Linear dimensions are measured in mm, mass in g. **b**, Technological variables. Centre lines show data median, boxes show first and third quartiles (the 25th and 75th percentiles), and the whiskers extend from the upper and lower hinge to the largest and smallest values no further than 1.5 times the interquartile range from the hinge. Data beyond the end of the whiskers are outlying points and are plotted individually.

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0793-8

Identifying the pathways required for coping behaviours associated with sustained pain

Tianwen Huang[1,2,8], Shing-Hong Lin[1,2,8], Nathalie M. Malewicz[3], Yan Zhang[1,4,5], Ying Zhang[1,6], Martyn Goulding[7], Robert H. LaMotte[3] & Qiufu Ma[1,2]*

Animals and humans display two types of response to noxious stimuli. The first includes reflexive defensive responses that prevent or limit injury; a well-known example of these responses is the quick withdrawal of one's hand upon touching a hot object. When the first-line response fails to prevent tissue damage (for example, a finger is burnt), the resulting pain invokes a second-line coping response—such as licking the injured area to soothe suffering. However, the underlying neural circuits that drive these two strings of behaviour remain poorly understood. Here we show in mice that spinal neurons marked by coexpression of TAC1^{Cre} and LBX1^{Flpo} drive coping responses associated with pain. Ablation of these spinal neurons led to the loss of both persistent licking and conditioned aversion evoked by stimuli (including skin pinching and burn injury) that—in humans—produce sustained pain, without affecting any of the reflexive defensive reactions that we tested. This selective indifference to sustained pain resembles the phenotype seen in humans with lesions of medial thalamic nuclei[1–3]. Consistently, spinal TAC1-lineage neurons are connected to medial thalamic nuclei by direct projections and via indirect routes through the superior lateral parabrachial nuclei. Furthermore, the anatomical and functional segregation observed at the spinal level also applies to primary sensory neurons. For example, in response to noxious mechanical stimuli, MRGPRD- and TRPV1-positive nociceptors are required to elicit reflexive and coping responses, respectively. Our study therefore reveals a fundamental subdivision within the cutaneous somatosensory system, and challenges the validity of using reflexive defensive responses to measure sustained pain.

The preprotachykinin 1 (*Tac1*) gene is expressed in spinal neurons that respond to noxious stimuli[4]. We used *Tac1^{cre}* knock-in mice to characterize spinal TAC1-lineage neurons. Crossing *Tac1^{cre}* mice with tdTomato reporter mice revealed that 45.3 ± 3.3% of TAC1^{Cre}–tdTomato^{+} neurons expressed *Tac1* mRNA persistently, and 83.9 ± 2.1% of *Tac1* mRNA^{+} neurons coexpressed tdTomato (Fig. 1a). Most TAC1^{Cre}–tdTomato^{+} neurons are excitatory (Extended Data Fig. 1a–c). The neurokinin receptor NK1R marks most ascending-projection neurons in lamina I (ref. [5]), 36.6 ± 4.6% of which were labelled by tdTomato (Fig. 1b). To assess ascending projections, we produced intersectional *Tac1^{Cdx2}-tdTomato* mice, in which Flpo is driven from the *Cdx2* gene locus and marks spinal neurons[6]; as such, all tdTomato^{+} fibres in the brain originate from spinal TAC1^{CDX2} neurons defined by developmental co-expression of TAC1^{Cre} and CDX2^{Flpo} (Extended Data Fig. 1d, e). As a control, we labelled all spinal ascending projection neurons by crossing *Cdx2^{Flpo}* mice with tdTomato reporter mice.

We first examined thalamic nuclei. The ventral posterolateral nuclei, which receive inputs from the spinal cord, are required for sensory discrimination and to process the unpleasantness evoked by transient, moderately noxious stimuli[7]. The medial thalamic complex is required to process the unpleasantness evoked by sustained, intensely noxious stimuli[1–3]. TAC1^{CDX2}-marked fibres rarely innervate the ventral posterolateral nuclei (Fig. 1c, Extended Data Fig. 1f), in contrast to the extensive innervation by CDX2^{Flpo}-marked fibres (Extended Data Fig. 1f). The medial thalamic complex is composed of: (i) the medial and lateral habenular nuclei, (ii) the paraventricular thalamic nucleus and (iii) the medial thalamic nuclei composed of the dorsal, central and ventral sub-nuclei (Fig. 1d, e). Whereas CDX2^{Flpo}-marked fibres innervated all of these midline nuclei (Extended Data Fig. 1g, h), TAC1^{CDX2}-marked fibres displayed selective innervations in paraventricular and medial thalamic nuclei (Fig. 1d), as well as the most-medial part of the lateral habenular nuclei (Fig. 1d, arrows). Thus, TAC1^{Cre} marks a subset of spinothalamic projection neurons that terminate predominantly within the medial thalamic pathways (Fig. 1f).

The pontine lateral parabrachial nuclei serve as another key relay station[8–11], one which is organized along the dorsoventral axis. The external lateral parabrachial nuclei (PBel) that are located in the most-ventral position are marked by the expression of the calcitonin-gene-related peptide (CGRP) and send neuronal projections to the amygdala; they are crucial for rapid defensive reactions to external threats[8]. The dorsoventral lateral parabrachial nuclei (PBdvl) are involved in behavioural thermoregulation[10], and the superior lateral parabrachial nuclei (PBsl) located in the most dorsal position are activated by painful stimuli (see below). We found that TAC1^{CDX2}-marked fibres pass through the area lateral to PBel and PBdvl (Fig. 1g, h), and terminate at PBsl (Fig. 1g). As a control, CDX2^{Flpo}-marked fibres innervated all subnuclei (Extended Data Fig. 1i). A selective synaptic connection between TAC1 neurons and PBsl was confirmed by using the presynaptic bouton marking technique[12], and by performing electrophysiological recordings following optogenetic stimulation of terminals derived from spinal TAC1-lineage neurons (Extended Data Fig. 2a–d). Furthermore, retrograde labelling confirms that TAC1^{Cre} marks subsets of spinoparabrachial and spinothalamic projection neurons (Extended Data Fig. 3a, b). Notably, PBsl sends projections to the medial thalamic nuclei but not to the amygdala (Extended Data Fig. 3c, d). Thus, spinal TAC1-lineage neurons include ascending projection neurons that are both directly and indirectly connected to medial thalamic nuclei.

To carry out functional studies, we used an intersectional genetic strategy[6,13,14] to express the human diphtheria toxin receptor DTR in spinal neurons defined by the coexpression of TAC1^{Cre} and LBX1^{Flpo} (Extended Data Fig. 4a) (hereafter referred to as TAC1^{LBX1}). Mice in which TAC1^{LBX1} neurons were ablated—and which additionally carried the tdTomato reporter allele—were generated following diphtheria toxin injections, which resulted in an 88% reduction of spinal tdTomato^{+} neurons (Fig. 2a). Ablation was also observed in trigeminal nuclei, but not dorsal root ganglia or other brain regions (Extended Data Fig. 4b). Ablation of TAC1^{LBX1} neurons did not affect sensorimotor coordination or responses to innocuous tactile stimuli (Extended Data Fig. 4c, d).

[1]Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA, USA. [2]Department of Neurobiology, Harvard Medical School, Boston, MA, USA. [3]Department of Anesthesiology, Yale University School of Medicine, New Haven, CT, USA. [4]Institute of Acupuncture and Moxibustion, Fudan Institutes of Integrative Medicine; Department of Integrative Medicine and Neurobiology, School of Basic Medical Science, Fudan University, Shanghai, China. [5]Cell Electrophysiology Laboratory, Wannan Medical College, Wuhu, China. [6]Neuroscience Research Institute and Department of Neurobiology, School of Basic Medical Sciences, Key Laboratory for Neuroscience, Ministry of Education/National Health and Family Planning Commission, Peking University, Beijing, China. [7]Molecular Neurobiology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA, USA. [8]These authors contributed equally: Tianwen Huang, Shing-Hong Lin. *e-mail: Qiufu_Ma@dfci.harvard.edu

© 2019 Springer Nature Limited. All rights reserved.

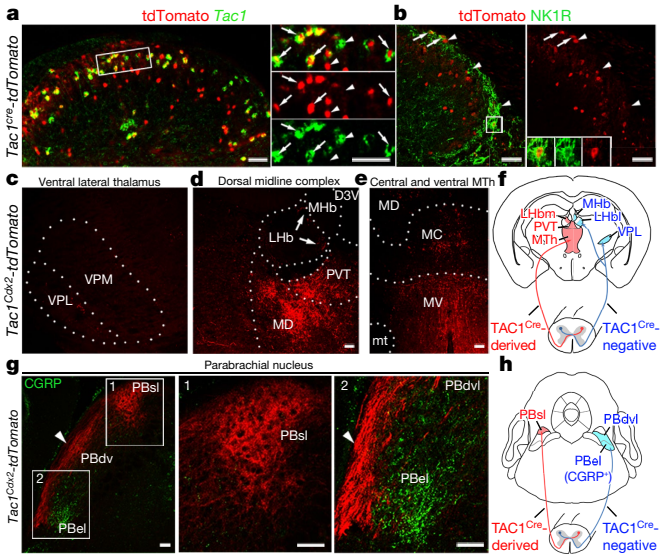


Fig. 1 | Spinal TAC1 neurons project to medial thalamic and superior lateral parabrachial nuclei. a, **b**, Representative sections from postnatal day 30 (P30) *Tac1*^{cre}*-tdTomato* mice ($n = 3$), showing tdTomato (red) with *Tac1* mRNA (green) or NK1R (green) in superficial dorsal spinal laminae. Arrows indicate co-localization and arrowheads indicate singular expression. Inset in **b** shows a NK1R^{+}tdTomato^{+} cell. **c**–**e**, Representative coronal thalamic sections (10 μm, bregma −1.70 mm) from P30 *Tac1*^{Cdx2}*-tdTomato* mice ($n = 3$). Arrows in **d** indicate the most-medial part of one of two lateral habenular nuclei (LHb). **f**, Schematic of thalamic projections. **g**, Left, a representative section through pontine lateral parabrachial nuclei (PBN) from P30 *Tac1*^{Cdx2}*-tdTomato* mice ($n = 3$), showing tdTomato (red) and CGRP immunostaining (green). Middle, note the innervations in PBsl. Right, arrowheads indicate fibres passing through the area lateral to PBdvl and CGRP^{+} PBel. **h**, Schematic of parabrachial projections. D3V, third ventricular, dorsal division; LHbl, lateral part of LHb; LHbm, medial part of LHb; MHb, medial habenular nuclei; mt, mammillothalamic tract; MTh, the medial thalamic nuclei including dorsal (MD), central (MC) and ventral (MV) sub-nuclei; PBdvl, dorsoventral lateral PBN; PBel, external lateral PBN; PBsl, superior lateral PBN; PVT, paraventricular thalamic nucleus; scp, superior cerebellar peduncle; VPL, ventral posterolateral thalamic nuclei; VPM, ventral posteromedial thalamic nuclei. Scale bars, 50 μm.

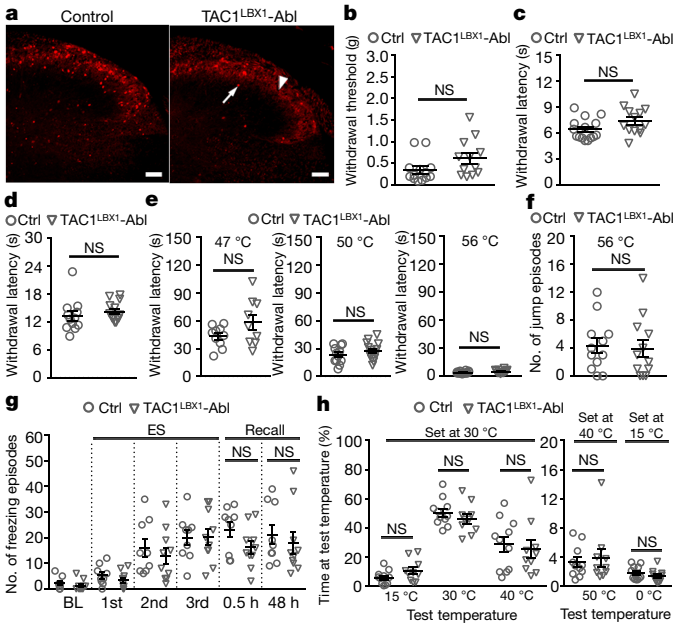


Fig. 2 | TAC1^{LBX1} neurons are dispensable for reflexive defensive reactions. a, Mice in which TAC1^{LBX1} neurons were ablated (TAC1^{LBX1}-Abl) showed a loss of spinal TAC1^{Cre}–tdTomato^{+} neurons (control mice, $n = 4$, neurons, 97 ± 7; TAC1^{LBX1}-Abl mice, $n = 4$, neurons, 11 ± 3; $P < 0.001$). Arrow, a remaining cell; arrowhead, processes from un-ablated primary afferents. Scale bars, 100 μm. **b**–**e**, Reflexive response tests. No significant (NS) differences in withdrawal responses to von Frey filament (**b**, $P = 0.09$), cold (**c**, $P = 0.07$), radiant heat (**d**, $P = 0.44$) or hot plate (**e**, $P = 0.11$, 0.28 and 0.12 for 47 °C, 50 °C and 56 °C, respectively) stimulus (control (ctrl), $n = 13$, 15, and 12 mice for **b**, **c** and **d**, respectively; **e**, $n = 10$, 12 and 14 mice for 47 °C, 50 °C and 56 °C, respectively; TAC1^{LBX1}-Abl, $n = 12$, 13 and 14 mice for **b**, **c** and **d**, respectively; **e**, $n = 9$, 15 and 12 mice for 47 °C, 50 °C and 56 °C, respectively). **f**, Comparable jumping evoked by 56-°C hot plate (control and TAC1^{LBX1}-Abl, $n = 12$ mice, $P = 0.80$). **g**, Foot shock test. Control ($n = 9$ mice) and TAC1^{LBX1}-Abl ($n = 10$ mice) groups showed no difference in freezing episodes by three repeated electrical stimulations (ES) (two-way ANOVA, $P = 0.86$), and no difference during recall phases (two-sided *t*-test: 0.5 h, $P = 0.10$; 48 h, $P = 0.58$; mean ± s.e.m.). BL, baseline. **h**, Two-plate preference tests. No difference in time spent at the test versus the set temperatures (control, $n = 11$ mice; TAC1^{LBX1}-Abl, $n = 10$ mice; $P = 0.10$, 0.33, 0.69, 0.67 and 0.28 for test temperatures at 15 °C, 30 °C, 40 °C, 50 °C and 0 °C, respectively). *P* values in **a**–**f**, **h** calculated by two-sided *t*-test. Data shown as mean ± s.e.m.

We next assessed reflexive defensive responses to noxious or threatening stimuli. Mice in which TAC1^{LBX1} neurons were ablated showed subtle but insignificant changes in withdrawal thresholds to punctate mechanical force evoked by von Frey filaments (Fig. 2b), in contrast to the abolition of such reflexes upon ablation of spinal somatostatin-lineage neurons[13]. The ablated mice also displayed normal, or subtle but insignificant changes in, withdrawal latencies to noxious cold (Fig. 2c) or heat (Fig. 2d, e) stimuli. Consistent with their sparse innervations to PBel and PBdvl, TAC1^{LBX1} neurons were dispensable for defensive reactions mediated by these nuclei. CGRP^{+} neurons in PBel are required to produce jumping when mice are confined to a 56-°C hot plate[8], and this escape response remained intact (Fig. 2f). PBel neurons also mediate freezing reactions following fearful events[8]. In an inescapable chamber, both control mice and mice in which TAC1^{LBX1} neurons were ablated displayed similar degrees of freezing immediately following repeated electric shocks (Fig. 2g). This was also true after the mice were brought back 30 min or 2 days later (Fig. 2g), which indicates that the development of conditioned fear memory to threatening events was normal. Neurons in PBdvl are necessary for behavioural thermoregulation[10], and control and ablated mice displayed indistinguishable temperature preferences (Fig. 2h). Thus, TAC1^{LBX1} neurons are largely dispensable for reflexive defensive responses to external threats.

We next assessed the behavioural responses that are evoked by prolonged noxious stimuli that should produce tissue damage and pain perception. We first placed individual mice in a hot or cold plate chamber until the cut-off time. On the 46–47-°C hot plate, wild-type mice showed temporally segregated behaviours, with paw lifting (often coupled with flinching) that preceded paw licking; the onset of licking correlated with Fos induction in TAC1^{Cre}-marked neurons (Extended Data Fig. 4e, f). Licking probably represents a form of coping behaviour that serves to soothe suffering invoked by pain, potentially via the activation of low-threshold mechanoreceptors that provide a gate control of pain[15,16]. Mice in which TAC1^{LBX1} neurons were ablated showed an abolition of the licking responses evoked by noxious thermal stimulation (Fig. 3a, b). The transient receptor potential channel TRPA1 serves as a sensor of tissue injury[17,18]. Intraplantar injection of mustard oil—a TRPA1 agonist that, in humans, produces sustained burning pain[19]—resulted in licking responses that lasted for 15 min in control mice, but virtually none in ablated mice (Fig. 3c). We further developed a model of the pain associated with hindpaw burn injury, which led to persistent licking throughout the 30-min post-injury period. Once again, this coping response was greatly reduced in mice in which TAC1^{LBX1} neurons were ablated (Fig. 3d). However, top-down execution of the licking motor behaviour per se is not impaired; ablated mice retained their licking responses to intraplantar capsaicin injection. This last

© 2019 Springer Nature Limited. All rights reserved.

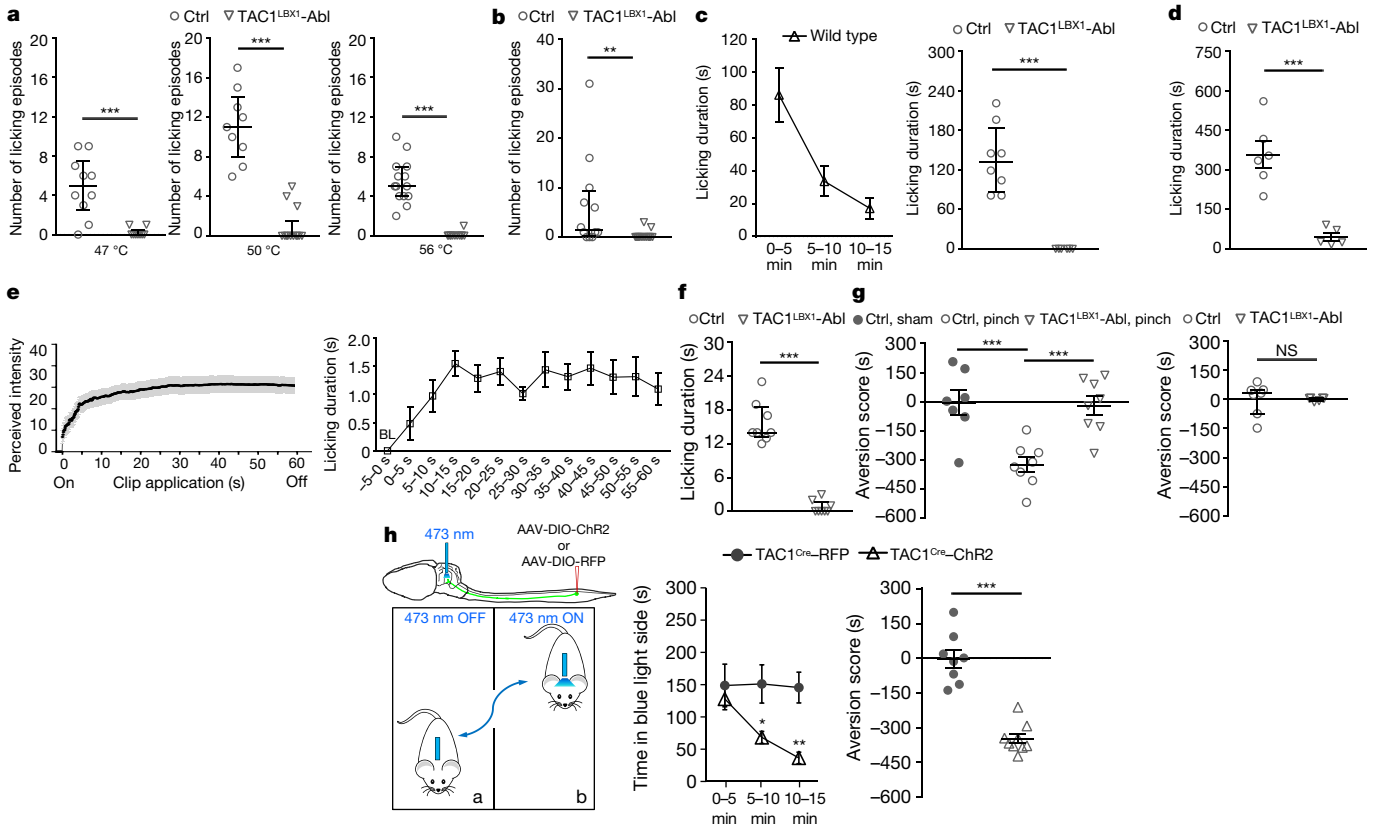


Fig. 3 | TAC1^{LBX1} neurons are required for noxious stimuli-evoked licking and CPA, and their activation drove CPA. a, **b**, TAC1^{LBX1}-Abl mice lost licking evoked by hot plate (**a**, control mice at 47 °C, $n = 10$; 50 °C, $n = 12$; and 56 °C, $n = 14$; TAC1^{LBX1}-Abl mice at 47 °C, $n = 9$; 50 °C, $n = 15$; and 56 °C, $n = 12$; $P < 0.001$) or cold plate (**b**, control mice, $n = 12$; TAC1^{LBX1}-Abl mice, $n = 15$; $P = 0.002$; χ^2 test for incidence of mice with licking: control mice, 9 out of 12; TAC1^{LBX1}-Abl mice, 2 out of 15; $\chi^2_{0.95,(1)} = 10.5$, $P = 0.001$). **c**, Mustard oil injection. Left, licking time course in wild-type mice ($n = 8$). The x axis is bins of minutes (0–5 min, 5–10 min, and 10–15 min; upper limits are not included in the bin). Right, loss of licking in TAC1^{LBX1}-Abl mice (control, $n = 8$ mice; TAC1^{LBX1}-Abl, $n = 6$ mice; $P < 0.001$). **d**, Reduced licking evoked by hindpaw burn injury (control, $n = 6$ mice; TAC1^{LBX1}-Abl, $n = 5$ mice; two-sided t-test, $P = 0.001$; mean ± s.e.m.). **e**, Skin pinching tests. Left, continuous pain ratings during a one-min period (human subjects, $n = 25$). Right, licking time course in wild-type mice ($n = 10$), counted every 5 s within a one-min period. **f**, Loss of pinch-evoked licking in TAC1^{LBX1}-Abl mice (control and TAC1^{LBX1}-Abl, $n = 8$ mice; $P < 0.001$). **g**, Left, hindpaw skin pinch-induced CPA in control male mice (sham handling group, $n = 7$ mice; pinched group, $n = 8$ mice; two-sided t-test, $P < 0.001$; mean ± s.e.m.), and CPA loss in TAC1^{LBX1}-Abl mice (pinched control and TAC1^{LBX1}-Abl, $n = 8$ mice; two-sided t-test, $P < 0.001$; mean ± s.e.m.). Right, radiant heat did not generate CPA (control and TAC1^{LBX1}-Abl, $n = 6$ mice; two-sided Mann–Whitney rank-sum test, $P = 0.234$, mean ± quartile). **h**, Left, intraspinal adeno-associated virus (AAV) injection and PBN implantation of the optic fibre in adult *Tac1^{cre}* mice. Blue light was on (30 Hz, 20-ms pulse width, 10 mW) whenever a mouse entered compartment b. Middle and right, optogenetic activation of terminals in PBN (control RFP mice, $n = 8$; ChR2 mice, $n = 9$) induced both acute avoidance (middle, two-way ANOVA followed by two-sided post hoc Bonferroni's t-test, $^*P = 0.012$, $^{**}P = 0.002$) and CPA 24 h later (right, two-sided t-test, $P < 0.001$; mean ± s.e.m.). The x axis of the middle panel is bins of minutes (0–5 min, 5–10 mins and 10–15 mins; upper limits are not included in the bin). **a**–**c**, **f**, Two-sided Mann–Whitney rank-sum test, data shown as mean ± quartile.

finding suggests the existence of pain pathways that are independent of TAC1^{LBX1} neurons (Extended Data Fig. 5).

Next, we assessed perceptions and behaviours invoked by sustained noxious mechanical stimuli. We performed human psychophysical studies that showed that pinching skin with an alligator clip evokes pain perception that has pricking, aching and burning components (Extended Data Fig. 6). Both the continuous pain ratings in humans and the licking behaviour in mice reach a peak within 10–15 s, and are maintained at peak levels throughout the remaining one-minute period (Fig. 3e). This persistent licking was virtually abolished in mice in which TAC1^{LBX1} neurons were ablated (Fig. 3f, Supplementary Videos 1, 2). Consistent with this finding, pinch-induced Fos expression in the dorsal spinal cord and in two brain nuclei innervated by spinal TAC1-lineage neurons—that is, PBsl and the medial thalamic complex—was greatly attenuated in ablated mice (Extended Data Fig. 7a–d). This selective loss of sustained coping responses indicative of affective pain mimics phenotypes seen in humans with lesions in medial thalamic nuclei[1,2], which is consistent with a direct and indirect connection of spinal TAC1 neurons to these nuclei (Fig. 1).

The differential effect of ablation of TAC1^{LBX1} neurons on distinct strings of behaviours was also apparent for stimuli that produce itch-related scratching. Short-lasting scratching evoked by external tactile stimulation may have evolved as a defensive reaction to remove insects or parasites that are touching the skin[14]. This defensive behaviour—evoked by light punctate stimulation onto the skin behind an ear—was preserved in mice in which TAC1^{LBX1} neurons were ablated (Extended Data Fig. 7e). By contrast, persistent scratching caused by inescapable intradermal exposure to pruritic compounds was greatly attenuated in ablated mice (Extended Data Fig. 7f); this is consistent with extensive innervations of spinal TAC1 neurons to PBsl, which are part of the parabrachial nuclei that process chemical itching[11]. Thus, for the stimuli tested so far (regardless of whether they are thermal, mechanical or pruritic), spinal TAC1^{LBX1} neurons are dispensable for reflexive defensive reactions to external threats, but necessary for coping behaviours that are directed towards sustained, inescapable injury or irritation of the skin.

Because sustained skin pinching produces pain and discomfort (Extended Data Fig. 6), we postulated that this pinching should

© 2019 Springer Nature Limited. All rights reserved.

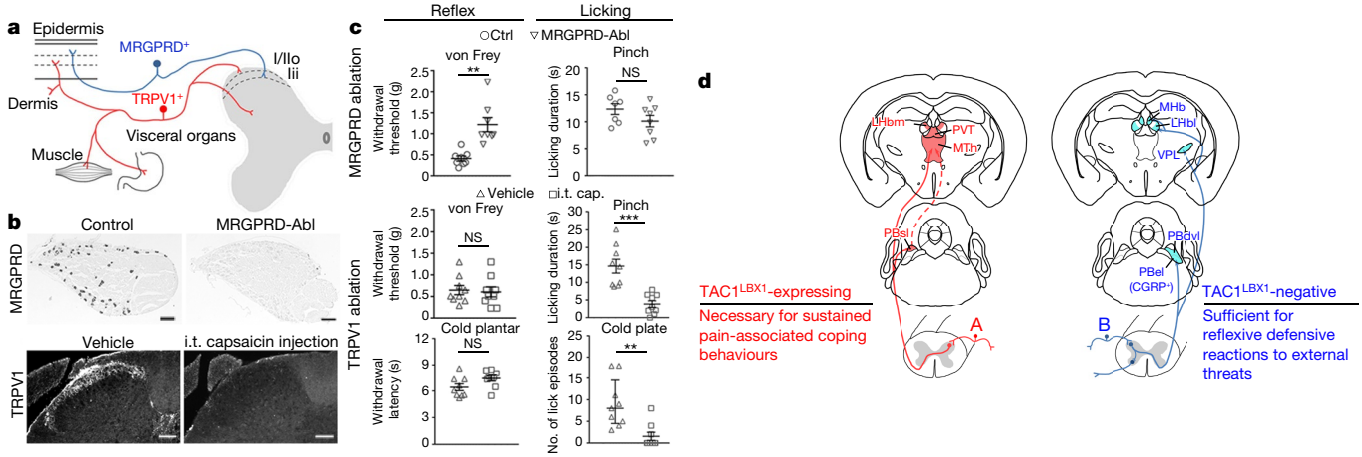


Fig. 4 | TRPV1$^+$, but not MRGPRD$^+$ neurons are required for noxious stimuli-evoked licking. a, MRGPRD$^+$ and TRPV1$^+$ neurons innervate distinct peripheral targets and spinal laminae. **b**, Top, representative in situ hybridization on dorsal root ganglia sections from three pairs of mice, indicating ablation of MRGPRD$^+$ neurons. Bottom, representative immunostaining images from three pairs of mice, showing ablation of TRPV1$^+$ terminals in superficial spinal lamianes following intrathecal (i.t.) capsaicin injection. Scale bars, 100 μm. **c**, Top, reduced reflexive responses to von Frey filaments in mice in which MRGPRD neurons were ablated (MRGPRD-Abl) (control and MRGPRD-Abl, $n = 8$ mice; two-sided t-test, $P = 0.002$), without affecting pinch-evoked licking (control, $n = 7$ mice; Abl, $n = 8$, two-sided t-test, $P = 0.150$). Middle and bottom, mice in which the TRPV1$^+$ terminal was ablated (TRPV1-Abl) (vehicle and capsaicin groups, $n = 9$ mice) showed reduced licking by pinch (two-sided t-test, $P < 0.001$) or cold-plate (Mann–Whitney rank-sum test for licking episode, $P = 0.002$; χ^2 test for incidence of mice with licking, control, 9 out of 9; TRPV1-Abl, 3 out of 9; $\chi^2_{0.95,(1)} = 9$, $P = 0.003$), without affecting withdrawal responses (two-sided t-test; von Frey, $P = 0.797$; cold plantar test, $P = 0.060$). **d**, Summary of two pathways for driving sustained pain-associated coping behaviours versus reflexive defensive reactions to external threats. A and B represent separate primary sensory neurons. A includes TRPV1$^+$ neurons required for licking evoked by pinch, noxious cold or heat, and skin burn injury, and B includes MRGPRD$^+$ neurons for reflexes evoked by von Frey filaments (for additional discussion, see Extended Data Fig. 10). Data are presented as mean ± s.e.m., except for the cold plate test (median ± quartile).

produce strong negative teaching signals that allow animals to learn to avoid such stimuli. To test this, we developed a pinch-evoked conditioned place aversion (CPA) assay (Extended Data Fig. 8a, b). After four training sessions, both male and female control littermates showed an aversion to the pinch-paired compartment and—as a control—sham handling did not produce CPA (Fig. 3g, Extended Data Fig. 8c). Mice in which TAC1^{LBX1} neurons were ablated were largely insensitive to this conditioning (Fig. 3g, Extended Data Fig. 8d). As a further comparison, repeated radiant heat stimuli—which elicited escapable withdrawal responses (Fig. 2d)—failed to produce CPA in either control or ablated mice (Fig. 3g). Thus, TAC1^{LBX1} neurons are necessary for conditioned learning and/or memory evoked by stimuli that produce sustained pain.

We next assessed whether activation of TAC1^{Cre}-expressing ascending terminals was sufficient to produce CPA. We used virus approaches to drive the expression of channelrhodopsin-2 fused with yellow fluorescent protein (ChR2–EYFP) or a red fluorescent protein (RFP) control, in adult lumbar spinal neurons that express TAC1^{Cre} (Fig. 3h, Extended Data Fig. 9a). Using a two-chamber avoidance assay[9], we found that optogenetic activation of TAC1^{Cre}-expressing terminals in PBsl induced robust avoidance during training sessions and post-training aversion to the stimulated chamber (Fig. 3h, Extended Data Fig. 9b, c). Stimulation of central terminals in the medial thalamic region produced similar consequences (Extended Data Fig. 9d–f). Thus, stimulation of TAC1^{Cre}-expressing ascending projection neurons produced sufficient negative teaching signals and aversive memory for CPA manifestation.

We next asked whether the anatomical and functional segregation observed at the spinal level applies to primary sensory neurons, by re-visiting two largely non-overlapping nociceptors marked by the expression of MRGPRD and TRPV1[20]. Anatomically, MRGPRD$^+$ and TRPV1$^+$ neurons innervate the skin epidermis[21] and the whole body[22,23], respectively (Fig. 4a). We used *Mrgprd*^{DTR} mice—in which human *DTR* is driven from the *Mrgprd* locus—to ablate adult MRGPRD$^+$ neurons, and intrathecal capsaicin injection to ablate the central terminals of TRPV1$^+$ nociceptors (Fig. 4b), as previously reported[20]. Ablation of MRGPRD$^+$ neurons caused an increase in withdrawal thresholds to the von Frey filament stimulation without changing pinch-evoked licking responses (Fig. 4c), which mimics phenotypes seen in mice with ablation of spinal VGLUT3^{Cre}-marked neurons[24]. Conversely, ablation of TRPV1$^+$ terminals did not affect withdrawal responses to von Frey filament and noxious cold stimulations, but caused a reduction in licking evoked by skin pinching, cold or hot plate stimulation, or skin burn injury (Fig. 4c, Extended Data Fig. 10a–c)—thereby mimicking phenotypes seen in mice in which TAC1^{LBX1} neurons have been ablated. We then found that pinch-induced Fos expression in TAC1^{Cre}-marked neurons was reduced upon ablation of TRPV1$^+$ terminals, which suggests a functional connection between these neurons (Extended Data Fig. 10d). It should be noted that, although TRPV1$^+$ nociceptors are required for both reflexes and licking evoked by noxious heat, separate subsets might be involved with each of these behaviours (Extended Data Fig. 10). Thus, by using coping rather than reflex assays, we have found the reverse of a previously held conclusion[20]: TRPV1$^+$ rather than MRGPRD$^+$ nociceptors are required to drive acute sustained mechanical pain. Our findings could explain the long-standing puzzle that, in response to skin pinching in humans, first-wave firing by polymodal nociceptors (probably including MRGPRD$^+$-like neurons) does not correlate with pain ratings[25]: pain ratings instead correlate with the firing of capsaicin-sensitive silent nociceptors that gain mechanical sensitivity during prolonged noxious stimuli[26,27].

In summary (Fig. 4d), our studies reveal a fundamental subdivision within the cutaneous somatosensory system that consists of separate pathways for driving reflexive defensive versus coping responses. These two strings of behaviours reflect exteroception (that is, sensitivity to external threats) versus interoception (that is, sensitivity to internal body injury), and serve to prevent or limit injury versus soothe suffering, respectively. Notably, the concurrent loss of licking or scratching evoked by inescapable noxious mechanical, heat, cold and pruritic stimuli suggests that spinal TAC1^{LBX1} neurons as a whole population—which should include both projection neurons and interneurons—have a general role in driving the affective component of sustained pain or itch in a manner that is independent of sensory modality. Our findings challenge the validity of using reflexive defensive responses to measure

© 2019 Springer Nature Limited. All rights reserved.

sustained pain, and their wide use for decades may have contributed to poor translation from preclinical studies to the development of new pain medicine[28,29], echoing a similar argument for the field of anxiety studies[30].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0793-8

Received: 2 January 2018; Accepted: 5 November 2018;
Published online 10 December 2018.

1. Mark, V. H., Ervin, F. R. & Yakovlev, P. I. Stereotactic thalamotomy III. The verification of anatomical lesion sites in the human thalamus. *Arch. Neurol.* **8**, 528–538 (1963).
2. Young, R. F. et al. Gamma knife thalamotomy for the treatment of persistent pain. *Stereotact. Funct. Neurosurg.* **64**, 172–181 (1995).
3. Price, D. D. Central neural mechanisms that interrelate sensory and affective dimensions of pain. *Mol. Interv.* **2**, 392–403, 339 (2002).
4. Gutierrez-Mecinas, M. et al. Preprotachykinin A is expressed by a distinct population of excitatory neurons in the mouse superficial spinal dorsal horn including cells that respond to noxious and pruritic stimuli. *Pain* **158**, 440–456 (2017).
5. Todd, A. J. Neuronal circuitry for pain processing in the dorsal horn. *Nat. Rev. Neurosci.* **11**, 823–836 (2010).
6. Bourane, S. et al. Identification of a spinal circuit for light touch and fine motor control. *Cell* **160**, 503–515 (2015).
7. Head, H. & Holmes, H. J. Sensory disturbances from cerebral lesions. *Brain* **34**, 102–154 (1911).
8. Han, S., Soleiman, M. T., Soden, M. E., Zweifel, L. S. & Palmiter, R. D. Elucidating an affective pain circuit that creates a threat memory. *Cell* **162**, 363–374 (2015).
9. Rodriguez, E. et al. A cranio-facial-specific monosynaptic circuit enables heightened affective pain. *Nat. Neurosci.* **20**, 1734–1743 (2017).
10. Yahiro, T., Kataoka, N., Nakamura, Y. & Nakamura, K. The lateral parabrachial nucleus, but not the thalamus, mediates thermosensory pathways for behavioural thermoregulation. *Sci. Rep.* **7**, 5031 (2017).
11. Mu, D. et al. A central neural circuit for itch sensation. *Science* **357**, 695–699 (2017).
12. Oh, S. W. et al. A mesoscale connectome of the mouse brain. *Nature* **508**, 207–214 (2014).
13. Duan, B. et al. Identification of spinal circuits transmitting and gating mechanical pain. *Cell* **159**, 1417–1432 (2014).
14. Bourane, S. et al. Gate control of mechanical itch by a subpopulation of spinal cord interneurons. *Science* **350**, 550–554 (2015).
15. Melzack, R. & Wall, P. D. Pain mechanisms: a new theory. *Science* **150**, 971–979 (1965).
16. Duan, B., Cheng, L. & Ma, Q. Spinal circuits transmitting mechanical pain and itch. *Neurosci. Bull.* **34**, 186–193 (2018).
17. Arenas, O. M. et al. Activation of planarian TRPA1 by reactive oxygen species reveals a conserved mechanism for animal nociception. *Nat. Neurosci.* **20**, 1686–1693 (2017).
18. Vandewauw, I. et al. A TRP channel trio mediates acute noxious heat sensing. *Nature* **555**, 662–666 (2018).
19. Ward, L., Wright, E. & McMahon, S. B. A comparison of the effects of noxious and innocuous counterstimuli on experimentally induced itch and pain. *Pain* **64**, 129–138 (1996).
20. Cavanaugh, D. J. et al. Distinct subsets of unmyelinated primary sensory fibers mediate behavioral responses to noxious thermal and mechanical stimuli. *Proc. Natl Acad. Sci. USA* **106**, 9075–9080 (2009).
21. Zylka, M. J., Rice, F. L. & Anderson, D. J. Topographically distinct epidermal nociceptive circuits revealed by axonal tracers targeted to *Mrgprd*. *Neuron* **45**, 17–25 (2005).
22. Cavanaugh, D. J. et al. Restriction of transient receptor potential vanilloid-1 to the peptidergic subset of primary afferent neurons follows its developmental downregulation in nonpeptidergic neurons. *J. Neurosci.* **31**, 10119–10127 (2011).
23. Malin, S. et al. TRPV1 and TRPA1 function and modulation are target tissue dependent. *J. Neurosci.* **31**, 10516–10528 (2011).
24. Cheng, L. et al. Identification of spinal circuits involved in touch-evoked dynamic mechanical pain. *Nat. Neurosci.* **20**, 804–814 (2017).
25. Adriaensen, H., Gybels, J., Handwerker, H. O. & Van Hees, J. Nociceptor discharges and sensations due to prolonged noxious mechanical stimulation—a paradox. *Hum. Neurobiol.* **3**, 53–58 (1984).
26. Schmidt, R., Schmelz, M., Torebjörk, H. E. & Handwerker, H. O. Mechano-insensitive nociceptors encode pain evoked by tonic pressure to human skin. *Neuroscience* **98**, 793–800 (2000).
27. Schmelz, M., Schmid, R., Handwerker, H. O. & Torebjörk, H. E. Encoding of burning pain from capsaicin-treated human skin in two categories of unmyelinated nerve fibres. *Brain* **123**, 560–571 (2000).
28. Cobos, E. J. & Portillo-Salido, E. 'Bedside-to-bench' behavioral outcomes in animal models of pain: beyond the evaluation of reflexes. *Curr. Neuropharmacol.* **11**, 560–591 (2013).
29. Borsook, D., Hargreaves, R., Bountra, C. & Porreca, F. Lost but making progress—where will new analgesic drugs come from? *Sci. Transl. Med.* **6**, 249sr3 (2014).
30. LeDoux, J. E. & Pine, D. S. Using neuroscience to help understand fear and anxiety: a two-system framework. *Am. J. Psychiatry* **173**, 1083–1093 (2016).

Acknowledgements We thank Z. J. Huang, M. J. Zylka, M. Hoon, S. M. Dymecki and the Allen Brain Institute/the Jackson Laboratory for genetically modified mice; A. I. Basbaum, A. V. Apkarian and C. J. Woolf for constructive discussions; and Y. Liu and Z. He for viral reagents. The work was supported by National Institutes of Health grants to Q.M. (R01 DE018025) and to Q.M. and M.G. (R01 NS086372), a Wellcome Trust grant to Q.M. (200183/Z/15/Z), and a Research Fellowship (326726541) from the German Research Foundation (DFG) to N.M.M.

Reviewer information *Nature* thanks F. Wang and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions T.H., S.-H.L., N.M.M., Yan Zhang and Ying Zhang performed the experiments and data analyses, Q.M., R.H.L. and M.G supervised the whole study, and T.H., S.-H.L., Q.M., R.H.L. and M.G. wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0793-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0793-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to Q.M.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

Mice. Mouse experiments were performed with protocols approved by the Institutional Animal Care and Use Committee. Mice were housed at room temperature with a 12-h/12-h light/dark cycle and had ad libitum access to standard laboratory mouse pellet food and water. *Tac1-IRES2-cre* knock-in mice (*Tac1*^{cre}, 021877), *Rosa26*$^{LSL\text{-}tdTomato}$ reporter mice (*Ai14*, 007908), *Rosa26*$^{LSL\text{-}FSF\text{-}tdTomato}$ reporter mice (*Ai65*, 021875) and *Rosa26*$^{LSL\text{-}FSF\text{-}CatCh}$ mice (*Ai80*, 025109) were acquired from Jackson Laboratory (JAX). The Flpo-dependent reporter strain was generated by crossing *Ai65* with *Meox2*^{cre} deleter strain (JAX, 003755) to remove the *loxP*-flanked STOP cassette (*LSL*). The intersectional GFP reporter line *Rosa26*$^{FSF\text{-}loxP\text{-}mCherry\text{-}STOP\text{-}loxP\text{-}EGFP}$ (*RC::FrePe*) was provided by S. M. Dymecki[31]. The generation of *Lbx1*^{Flpo}, *Tau*$^{ds\text{-}DTR}$ and *Cdx2*^{Flpo} mice has previously been described[6,13,32]. *Mrgprd*^{DTR} mice, produced in the D. J. Anderson laboratory[20], were provided by M. Hoon.

To generate mice in which TAC1^{LBX1} neurons were ablated, 6–10-week-old male and female *Tac1*$^{cre/+}$;*Lbx1*$^{Flpo/+}$;*Tau*$^{ds\text{-}DTR/+}$ mice (additionally carrying the *Ai14* reporter allele) were injected intraperitoneally with diphtheria toxin (50 μg/kg, Sigma-Aldrich, D0564) twice with a 72-h interval. Behavioural and histochemical analyses were performed 30 days later. Littermates that lacked either the *Lbx1*^{Flpo} or *Tac1*^{cre} allele but received the same diphtheria toxin injections were used as a control. To ablate Mrgprd^{+} neurons, we performed 5 consecutive daily intraperitoneal injections of diphtheria toxin (20 μg/kg) in *Mrgprd*$^{DTR/+}$ mice and their control littermates.

To characterize TAC1^{Cre}–tdTomato^{+} neurons, we analysed 18–25 lumbar spinal cord sections from three P30 *Tac1*^{cre};*Ai14* mice. To test ablation efficiency, we used 4 pairs of 12-week-old control and ablated mice. For each behavioural analysis, 6–15 pairs of 10–14-week-old ablated and control mice—including males and females—were used. Mice were randomly assigned into treatment groups and behaviour responses were measured in a blinded manner.

Intrathecal capsaicin injection. Eight-week-old 129 wild-type mice (JAX, 002448) were anaesthetized with 2% isoflurane and injected with 5 μl capsaicin (intrathecal, 10 μg, Sigma, M2028) or vehicle (10% ethanol and 10% Tween-80 in saline) at the pelvic girdle level. Behavioural tests were performed between 7 and 14 days later.

Fluorogold retrograde labelling. Adult *Tac1*^{cre}-*tdTomato* mice were anaesthetized by intraperitoneal injection of a ketamine–xylazine mixture (87.5 and 12.5 mg per kg, respectively). The head was mounted on the stereotaxic frame (Stoelting). After surgical exposure of the brain surface, 0.3 μl fluorogold (fluorochrome, 2% in sterilized water) was injected into the medial thalamic complex (anterior–posterior (AP), −1.58 mm; medial–lateral (ML), 0 mm; dorsal–ventral (DV), −2.5 mm from brain surface) or lateral parabrachial nucleus (AP, −5.1 mm; ML, −1.3 mm; DV, −2.3 mm from brain surface) via a glass micropipette driven by a picospritzer. The animals recovered for seven days before tissue collection.

Electrophysiological recording. The brains from 6-week-old *Tac1*^{Cdx2}-*CatCh* mice—the generation of which is described in Extended Data Fig. 2—were freshly dissected and coronally sectioned using a Leica-VT1000s vibratome at 400-μm thickness in ice-cold modified artificial cerebrospinal fluid, as previously reported[24]. Brain slices that covered parabrachial nuclei (bregma −5.20 ± 0.20 mm) were incubated in the recording solution as previously reported[24], for at least 1 h in room temperature. The brain slice was then transferred into a recording chamber and perfused with oxygenated recording solution as previously reported[24]. The location of PBsl and PBel subnuclei were visually identified under bright field and neurons within the target area were randomly picked and patched extracellularly. The 473-nm blue light (0.2 Hz, 20-ms wave width, 10 mW) was delivered and the responses were recorded at voltage clamp mode (holding membrane potential at −70 mV) and then at current clamp mode.

Intraspinal AAV viral injection. The Cre-dependent AAV plasmids phSyn-1(S)-FLEX-tdTomato-T2A-SypEGFP-WPRE (51509) and pAAV-EF1α-DIO-hChR2(H134R)-EYFP-WPRE-HGHpA (AAV-DIO-ChR2, 20298) were acquired from Addgene. The hChR2(H134R)–EYFP fragment was replaced by a RFP cassette to generate plasmid pAAV-EF1α-DIO-RFP-WPRE-HGHpA. These three plasmids were then packed in the AAV2/8 serotype at the Boston Children's Hospital Viral Core, at the titres of 1.06 ×10^{15} gc/ml, 5.776 × 10^{13} gc/ml and 4.520 × 10^{13} gc/ml, respectively. Adult *Tac1*^{cre} mice were anaesthetized by ketamine–xylazine (see 'Fluorogold retrograde labelling'). Lumbar vertebrae were exposed by laminectomy, and the AAVs were infused bilaterally into the lateral border of the grey and white matter of the dorsal horn (3 spots for each side, 0.3 μl/spot), at 300 μm from dorsal surface and with 500-μm interspaces. AAVs were delivered by picospritzer via a glass micropipette. The tip was left in the spinal cord for additional 5 min before being pulled out. After a three-week recovery, mice were used for optogenetic studies (see below) or euthanized for histochemical studies.

Stereotaxic optic fibre implantation. Under anaesthetic conditions, the mouse head was mounted on the stereotaxic frame as described above. After brain exposure, an optical fibre (material: Ceram; ferrule OD 2.5 mm; fibre core, 400 μm; fibre length, 3.0 mm; ThorLabs) was implanted above the PBsl (AP, −5.1 mm; ML, −1.3 mm; DV, −2.3 mm from brain surface) or above the medial thalamic complex (AP, −1.58 mm; ML, 0 mm; DV, −2.5 mm from brain surface). The optic fibres were secured by mounting dental cement on the skull. After a one-week recovery, mice were subjected to behavioural studies, or Fos induction analyses following optogenetic stimulations.

Histology. Mice were euthanized by CO_2 and then perfused with 4% paraformaldehyde prepared in 1× PBS, pH7.4. The brain, spinal cord and lumbar dorsal root ganglia were dissected and processed (25-μm thickness for brain sections, 16-μm for spinal cord and dorsal root ganglia sections) for immunofluorescent staining as previously described[13,24].

To investigate pinch- or blue-light-induced Fos expression, mice were euthanized 120 min after stimulation, and post-fixed brains were sectioned with vibratome at 75-μm thickness. Free-floating sections were performed to detect Fos^{+} cells using the VECTASTAIN Elite ABC Kits (Vector Labs, PK-6101). For co-staining of ChR2–EYFP with CGRP in PBN, after visualizing ChR2–EYFP^{+} fibres using nickel–DAB solution (dark blue), brain slices were incubated with mouse anti-CGRP for 16 h at 4 °C. Sections were then washed with PBS with Tween 20 and incubated with HRP-conjugated goat-anti-mouse secondary antibody for 1 h at room temperature. CGRP^{+} signals (brown) were visualized with DAB in PBS containing 0.03% H_2O_2. Colour images were taken using a Leica DMLB microscope.

Primary antibodies were rabbit anti-Fos (Millipore, ABE457, 1:1,000 for immunofluorescent and 1:5,000 for DAB staining), rabbit anti-CGRP (Millipore, PC205, 1:1,000), mouse anti-CGRP (Sigma, C7113, 1:1,000), rabbit anti-GFP (Invitrogen, A11122, 1:1,000), rabbit anti-NK1R (Sigma, S8305, 1:1,000), rabbit anti-TRPV1 (Alomone Labs, ACC-030, 1:1,000). Secondary antibodies: goat anti-rabbit IgG-Alexa 488 (Invitrogen, A11034, 1:500), biotinylated goat-anti-rabbit IgG (Vector Labs, BA-1000, 1:1,000) and HRP-conjugated goat-anti-mouse IgG (Bio-Rad, 1706516, 1:1,000).

In situ hybridization combined with immunohistochemistry procedures were performed as previously described[33,34].

Behavioural tests. For all behavioural tests, animals were habituated for 30 min per day consecutively for 3 days before testing. The experimenters were blinded to genotypes and treatments. The subsequent statistical analyses included all data points; no methods were used to predetermine sample sizes. Littermates were used as control.

The following behavioural assays were performed as previously described[6,13,24,35–37]: rotarod, light touch (brush), radiant heat (Hargreave's), cold plantar test (dry ice), hot plate, cold plate, von Frey filament test, two-plate temperature preference test, and touch- or chemical-compound-induced scratching test. In hot and cold plate tests, hindpaw lifting coupled with flinching was used to determine withdrawal latencies, followed by determining the duration of licking. Different cutoffs were set for each hot or cold plate test: 4 min for 46 °C, 3 min for 47 °C, 1 min for 50 °C, 30 s for 56 °C and 5 min for 0 °C. For Fos induction by different hot plate temperatures (see below), the cutoff was 3 min for both 46 °C and 50 °C.

For touch-evoked scratching behaviour, each mouse was lightly anaesthetized by 2% isoflurane and the hairs behind the ear were shaved. After recovering from anaesthesia, animals were placed into the test chamber for a 1-h habituation. A von Frey filament (0.07 g) was used to poke the shaved skin area. The incidence of scratching responses in ten trials was determined for each mouse.

Mustard oil test was performed by intraplantar injection to one hindpaw (Sigma, 377430-5G, 0.15% in saline, 20 μl/animal). The capsaicin test was performed similarly (Sigma, M2028, either 3 μg or 0.03 μg/animal in 10 μl saline containing 10% ethanol and 10% Tween-80). Animals were video-recorded for 15 min (mustard oil) or 5 min (capsaicin) to determine licking durations.

The pain test for skin burn injury was performed by immersing the distal half of the right hindpaw of the mouse into a 60-°C water bath for 30 s under deep anaesthesia condition (3% isoflurane inhalation via a precise vaporizer). Each mouse was placed in an observation chamber (7.5-cm long × 7.5-cm wide × 7.5-cm high) to recover from anaesthesia (which took about 5 min), and then video-recorded for 30 min to determine licking durations. The mice were euthanized immediately, because pilot studies had shown that necrosis would have developed within 3 days.

For pinch assay, each mouse was confined in a plexiglass chamber (10-cm long × 10-cm wide × 12-cm high) placed onto a glass, allowing video recording from the bottom. An alligator clip (Amazon, 'Generic Micro Steel Toothless Alligator Test Clips 5AMP') producing 340*g* force (see 'Skin pinch study in human subjects' for force calculation) was applied to the ventral skin surface between the footpad and the heel (Extended Data Fig. 8a). The animal was placed back into the chamber and video-recorded for 60 s to determine licking duration.

To measure pinch-evoked negative valence, we developed a CPA assay. The CPA chamber (30-cm long × 30-cm wide × 20-cm high) included compartments a and b. In compartment a, the wall was black plexiglass and the floor was made up of a set of stainless steel bars (3 mm in diameter, 1-cm interval between bars);

© 2019 Springer Nature Limited. All rights reserved.

in compartment b, the wall was decorated with black and white vertical strips of 1-cm width, and the floor was made up of a piece of stainless steel with 8-mm-diameter holes. The procedure included three sessions: baseline measurement (day 1), training (days 2–5) and test (day 6). On every training day, in the morning each mouse was confined in compartment a, and was grabbed three times with a 5-min interval between each handling session, without pinch stimulation; in the afternoon, each mouse was confined to compartment b, and received three trials of hindpaw skin pinching (1 min for each trial, with 4-min interval between each trial). For baseline and test-day measurement, each mouse had free access to both compartments for 15 min, and the time spent in compartment b was determined. A set of wild-type littermates was used as sham control; these mice received grabbing without pinching in both chambers. For radiant-heat-evoked CPA measurement, each training session contained 15 trials of radiant heat stimulation to the hindpaw. In our radiant heat test, the average withdrawal latency for wild-type mice was ~12 s; as such, the total stimulation time ($12 \times 15 = 180$ s) matched the total duration of skin pinching for the pinch-evoked CPA assay.

In foot shock test, animals were habituated for three days. On test day, each mouse was allowed to freely explore a customized foot shock chamber for 5 min, followed by three consecutive electric foot shocks (0.5 mA, 2-s duration, 3-min interval). The mice were returned to their home cage. Both 30 min and 48 h later, we then performed recall tests by placing mice back into the shock-paired chamber. The mice were video-recorded, and freezing episodes were defined as complete absence of movement except for animal breathing, which was automatically judged and analysed by ANY-maze behaviour tracking software using default settings (freezing 'on' threshold score = 30, freezing 'off' threshold score = 40, minimum freeze duration = 250 ms). The electric shock was delivered by a stimulator (Model 3800, A-M Systems) and an isolator (Model #3820, A-M Systems).

Two-trial avoidance task with optogenetic activation. *Tac1*^{cre} mice with intraspinal injection of either AAV-DIO-ChR2 or AAV-DIO-RFP were used for the PBN optogenetic activation test. A four-day training programme was performed. The apparatus was similar as that used for the pinch-induced CPA, with small modification of texture cues: in compartment a, the floor was composed of thinner stainless steel bars (1 mm in diameter, 2-mm interval); in compartment b, the floor was made up of a piece of wire mesh with grid of 5 mm × 5 mm. Mice had no initial preference for either one of the two compartments (compartment a, 436 ± 26 s versus compartment b: 464 ± 26 s, $n = 9$ mice, two-sided paired *t*-test, $P = 0.604$). On day 1 and day 4, each mouse was allowed to freely explore the chamber for 15 min. Avoidance training sessions (15 min/trial) were performed on day 2 (trial 1) and day 3 (trial 2). When a mouse walked into and stayed in the paired compartment, the 473-nm blue light (30 Hz, 20-ms pulse width, 10 mW, Opto Engine, Laser Model PSU-III-LED) was delivered to PBN. The laser was turned off immediately after the mouse left the paired compartment with its four paws. Mice were video-recorded and the times spent in the blue-light-paired compartment were calculated. The aversion score was calculated by the reduction of time spent in blue-light-paired compartment on day 4 (t_2) in comparison with day 1 (t_1), or during the trials.

The *Tac1*^{Cdx2}*-CatCh* and control *Tac1*^{Cdx2}*-GFP* mice were used for optogenetic activation of terminals derived from spinal TAC1^{Cre}-marked neurons in the medial thalamic complex. The same programme was performed as described above.

Fos induction. To study heat-evoked Fos expression, each mouse was placed on a 46-°C or 50-°C hot plate for 3 min. To study the pinch-evoked Fos expression in the spinal cord, each mouse was first habituated as it did for the pinch behavioural assay (see 'Behavioural tests'), and 3 trials of a 30-s hindpaw skin pinch was applied, with a 5-min interval between pinches.

To study pinch-evoked Fos expression in the brain, we had to minimize background Fos expression. To do this, each mouse was housed alone for three days; animals were subjected to gentle grabbing and holding for 10 s, 5 times every day in their home cages, and the mouse was also placed into a new empty cage individually for 120 min, before returning to the original home cage. On the test day, each mouse was habituated in the aforementioned empty cage for 30 min, and a single trial of a 180-s skin pinch was applied to left hindpaw.

To study blue-light-induced Fos expression, each mouse expressing ChR2, RFP, CatCh or GFP was habituated in the CPA apparatus for 4 h, and then received a 15-min (20-s light on, with 40-s light off interval) blue-light stimulation (30 Hz, 20-ms pulse width, 10 mW).

After pinch or blue-light stimulation, animals were kept in the same apparatus for 120 min, and spinal and brain tissues were then processed as described above.

Skin pinch study in human subjects. The protocols for the recruitment and testing of human subjects were approved by the Yale University Human Investigative Committee. Twelve healthy women and thirteen healthy men gave their written, informed consent to participate in the study. The aim was to measure the time course, magnitude and quality of pain sensation reported by humans when pinched by the alligator clip used in similar fashion to test pain-like behaviour in mice. First, the force applied by the clip was measured in the Yale Medical School machine shop (by A. DeSimone): With the bottom jaw of the clip fixed in a milling machine, a wire attached at one end to the upper jaw (strengthened to prevent bending) and the other end of the wire attached to a digital force-measuring device, the minimal lifting force required to achieve an opening or gap of 0.1 mm between the jaws was 340*g* (approximately the width of the skin fold between the jaws when the clip was applied to the hindpaw skin of a mouse). A force of 440*g* was required for an opening-gap of 1 mm (approximately the width of the skin fold between the jaws when the clip was applied to the forearm skin of the human subject with 3–4-mm length of skin on each side). Next, we measured nociceptive sensations. The 25 subjects were instructed to use the generalized labelled magnitude scale (GLMS)[38,39] to make continuous ratings of the maximal perceived intensity of pain evoked during a one-minute pinch produced by the alligator clip, as applied to a fold of skin on the mid-volar forearm. The GLMS was displayed on a computer screen and consisted of a vertical, thermometer-like scale with labels spaced quasi-logarithmically along its length—from 'no sensation' at the bottom, through 'barely detectable', 'weak', 'moderate', 'strong' and 'very strong', to 'strongest imaginable sensation of any kind' at the top. The subject was told to rate continuously the maximal perceived intensity of pain regardless of its sensory quality by moving a cursor along the scale by means of a computer mouse. Subjects ratings were recorded continuously every 100 ms of the 1-min pinch stimulation period, and evaluated and displayed as pain rating over time. The area under the curve for the whole 1-min pinch stimulation for each subjects' rating was calculated using GraphPad Prism 7. After the clip was removed, the subject was asked to rate the maximal perceived intensity of each of four aversive qualities of cutaneous sensation associated with the pain they had just experienced (that is, itch, pricking or stinging, burning and aching). Then, they were asked to rate the intensity of any feeling of discomfort associated with this maximal sensation. The advantage of the GLMS is that it allows the perceived intensities of different cutaneous qualities to be compared on a common scale[40–42]. Before testing, the subjects were told that they might or might not experience one or more of these sensory qualities.

Statistics. Statistical analyses were performed by using SigmaStat 3.5 and GraphPad Prism 7 software. All datasets were tested for normality for *t*-test, and if the normality test failed, the Mann–Whitney rank-sum test was used. Results are expressed as mean ± s.e.m. or median ± quartile. $P < 0.05$ is considered as significant. For the evaluation of ablation efficiency, and pinch- or blue-light-evoked Fos induction, data were subjected to a two-sided Student's *t*-test. For pinch-evoked Fos in LHb and PBN, data were assessed by two-way ANOVA followed by a post hoc Holm–Sidak's *t*-test. The χ^2 test was applied to determine the incidence of ventral posterolateral nucleus innervation detected in thalamic sections, the incidence of licking evoked by the cold plate and the incidence of recorded neurons showing activation by optogenetic stimulation. Behaviour data were analysed by using two-sided Student's *t*-test, Mann–Whitney rank-sum test or two-way ANOVA followed by a post hoc Bonferroni's *t*-test. Human psychophysical data were analysed with a two-sided Student's *t*-test or Mann–Whitney rank-sum test. The continuous pain ratings over time were analysed by two-way repeated-measures ANOVA (gender × time). No statistical methods were used to predetermine sample sizes.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

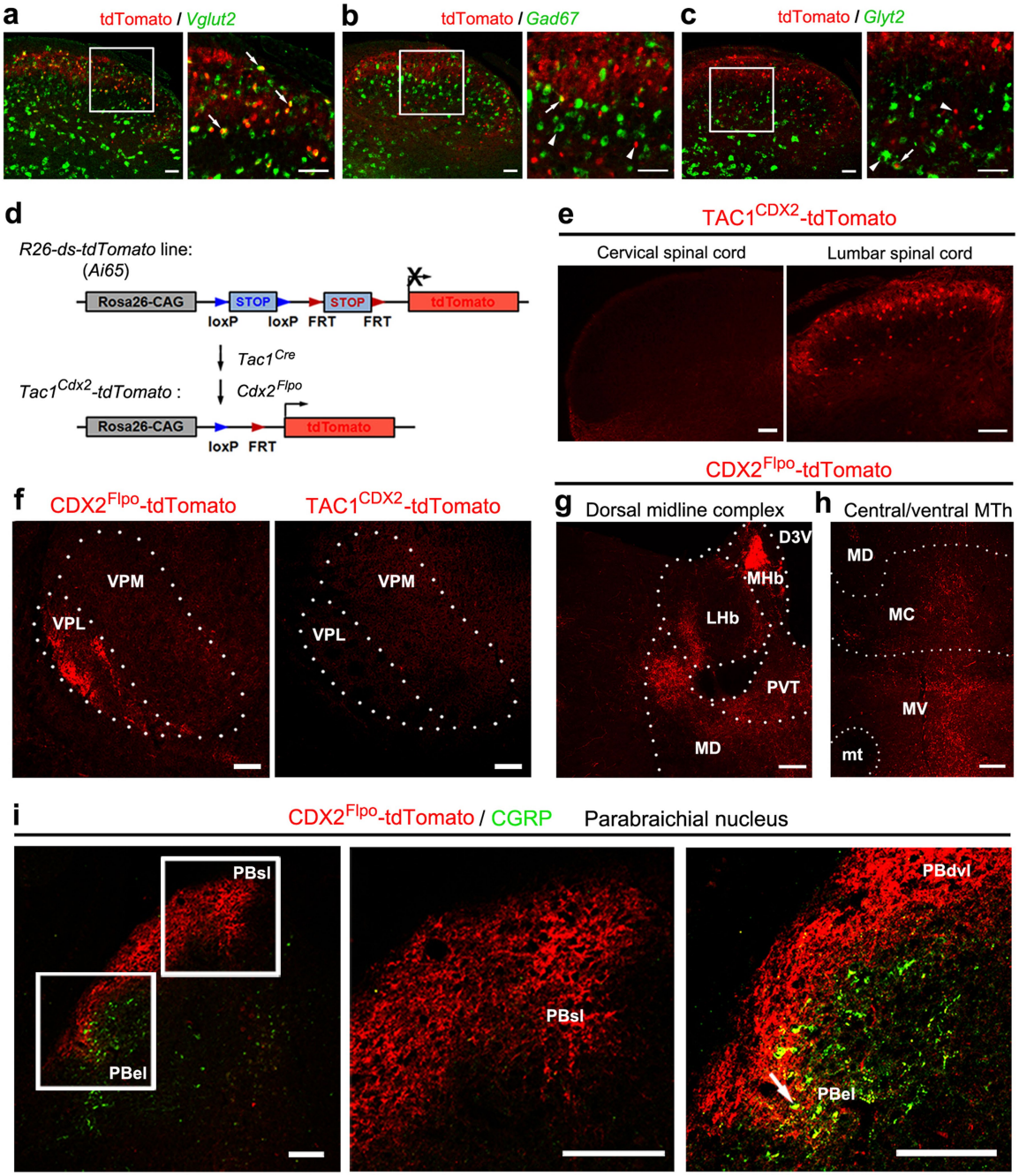
All data are contained in the main text or Supplementary Materials or are available from the corresponding author upon reasonable request. The anterograde tracing data are from the Allen Brain Atlas website (http://connectivity.brain-map.org/).

31. Niederkofler, V. et al. Identification of serotonergic neuronal modules that affect aggressive behavior. *Cell Rep.* **17**, 1934–1949 (2016).
32. Britz, O. et al. A genetically defined asymmetry underlies the inhibitory control of flexor–extensor locomotor movements. *eLife* **4**, e04718 (2015).
33. Liu, Y. et al. VGLUT2-dependent glutamate release from nociceptors is required to sense pain and suppress itch. *Neuron* **68**, 543–556 (2010).
34. Lou, S., Duan, B., Vong, L., Lowell, B. B. & Ma, Q. Runx1 controls terminal morphology and mechanosensitivity of VGLUT3-expressing C-mechanoreceptors. *J. Neurosci.* **33**, 870–882 (2013).
35. Knowlton, W. M., Bifolck-Fisher, A., Bautista, D. M. & McKemy, D. D. TRPM8, but not TRPA1, is required for neural and behavioral responses to acute noxious cold temperatures and cold-mimetics in vivo. *Pain* **150**, 340–350 (2010).
36. Brenner, D. S., Golden, J. P. & Gereau, R. W., IV. A novel behavioral assay for measuring cold sensation in mice. *PLoS ONE* **7**, e39765 (2012).
37. Pogorzala, L. A., Mishra, S. K. & Hoon, M. A. The cellular code for mammalian thermosensation. *J. Neurosci.* **33**, 5533–5541 (2013).
38. Green, B. G. et al. Evaluating the 'Labeled Magnitude Scale' for measuring sensations of taste and smell. *Chem. Senses* **21**, 323–334 (1996).
39. Bartoshuk, L. M. et al. Valid across-group comparisons with labeled scales: the gLMS versus magnitude matching. *Physiol. Behav.* **82**, 109–114 (2004).
40. Green, B. G. & Schoen, K. L. Thermal and nociceptive sensations from menthol and their suppression by dynamic contact. *Behav. Brain Res.* **176**, 284–291 (2007).

© 2019 Springer Nature Limited. All rights reserved.

41. LaMotte, R. H., Shimada, S. G., Green, B. G. & Zelterman, D. Pruritic and nociceptive sensations and dysesthesias from a spicule of cowhage. *J. Neurophysiol*. **101**, 1430–1443 (2009).
42. Sikand, P., Shimada, S. G., Green, B. G. & LaMotte, R. H. Sensory responses to injection and punctate application of capsaicin and histamine to the skin. *Pain* **152**, 2485–2494 (2011).
43. Kleinlogel, S. et al. Ultra light-sensitive and fast neuronal activation with the Ca^{2+}-permeable channelrhodopsin CatCh. *Nat. Neurosci*. **14**, 513–518 (2011).
44. Kwan, K. Y. et al. TRPA1 contributes to cold, mechanical, and chemical nociception but is not essential for hair-cell transduction. *Neuron* **50**, 277–289 (2006).
45. Story, G. M. et al. ANKTM1, a TRP-like channel expressed in nociceptive neurons, is activated by cold temperatures. *Cell* **112**, 819–829 (2003).
46. Bautista, D. M. et al. TRPA1 mediates the inflammatory actions of environmental irritants and proalgesic agents. *Cell* **124**, 1269–1282 (2006).
47. Wang, S. et al. Phospholipase C and protein kinase A mediate bradykinin sensitization of TRPA1: a molecular mechanism of inflammatory pain. *Brain* **131**, 1241–1251 (2008).
48. Bartley, E. J. & Fillingim, R. B. Sex differences in pain: a brief review of clinical and experimental findings. *Br. J. Anaesth*. **111**, 52–58 (2013).
49. Doehring, A. et al. Effect sizes in experimental pain produced by gender, genetic variants and sensitization procedures. *PLoS ONE* **6**, e17724 (2011).
50. Giesler, G. J. Jr, Yezierski, R. P., Gerhart, K. D. & Willis, W. D. Spinothalamic tract neurons that project to medial and/or lateral thalamic nuclei: evidence for a physiologically novel population of spinal cord neurons. *J. Neurophysiol*. **46**, 1285–1308 (1981).
51. Benabid, A. L. & Jeaugey, L. Cells of the rat lateral habenula respond to high-threshold somatosensory inputs. *Neurosci. Lett*. **96**, 289–294 (1989).
52. Abdel Samad, O. et al. Characterization of two Runx1-dependent nociceptor differentiation programs necessary for inflammatory versus neuropathic pain. *Mol. Pain* **6**, 45 (2010).
53. Kiasalari, Z. et al. Identification of perineal sensory neurons activated by innocuous heat. *J. Comp. Neurol*. **518**, 137–162 (2010).
54. Chen, C. L. et al. Runx1 determines nociceptive sensory neuron phenotype and is required for thermal and neuropathic pain. *Neuron* **49**, 365–377 (2006).

© 2019 Springer Nature Limited. All rights reserved.

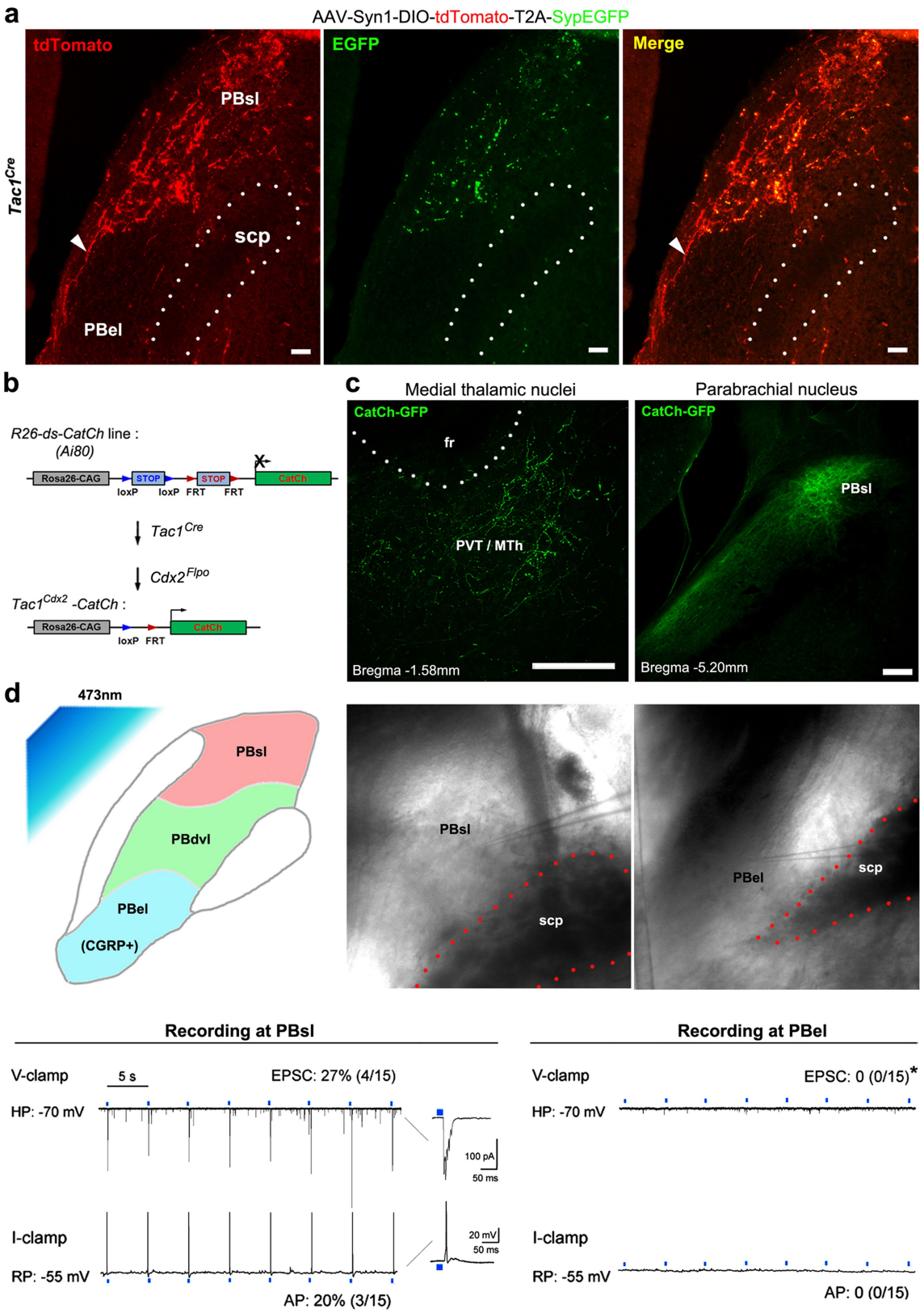


Extended Data Fig. 1 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Fig. 1 | Neurotransmitter phenotypes and central projection by spinal TAC1^{Cre+} neurons. a–c, Lumbar spinal sections from P30 *Tac1^{cre}-tdTomato* mice ($n = 3$), in which spinal neurons with developmental expression of TAC1^{Cre} were labelled by tdTomato, showing double staining of tdTomato signals (red) and the mRNA of the excitatory neuronal marker VGLUT2 (**a**, green), or inhibitory neuronal markers GAD67 (**b**, green) and GlyT2 (**c**, green) detected by in situ hybridization[13]. Right panels represent a higher magnification of the boxed areas. Arrows show co-localization, and arrowheads show singular expression. Quantification of neurotransmitter phenotypes of spinal TAC1^{Cre}–tdTomato^{+} neurons: 91.2 ± 0.7% (± s.e.m.) are VGLUT2^{+} excitatory neurons, 6.1 ± 1.1% (± s.e.m.) are GAD67^{+} GABAergic inhibitory neurons and 4.5 ± 0.9% (± s.e.m.) are GlyT2^{+} glycinergic inhibitory neurons. **d**, Intersectional genetic strategy for driving tdTomato expression in spinal TAC1^{CDX2} neurons defined by co-expression of TAC1^{Cre} and CDX2^{Flpo}. It had previously been reported that CDX2^{Flpo} drives reporter expression from the cervical spinal cord all the way to the most-caudal spinal cord[6]. By crossing *Tac1^{cre}* mice and *Cdx2^{Flpo}* mice with intersectional *Ai65* reporter mice, only spinal neurons with developmental co-expression of TAC1^{Cre} and CDX2^{Flpo} drove tdTomato expression; these are referred to as *Tac1^{Cdx2}-tdTomato* mice. **e**, Representative sections through the spinal cord of *Tac1^{Cdx2}-tdTomato* mice ($n = 3$), showing that tdTomato^{+} neurons are not detected in the most-rostral cervical levels or in the brain (data not shown), but are detected in the lumbar levels. **f**, Representative coronal sections (25-μm thick, prepared by cryostat; in comparison with Fig. 1c, which was 100-μm thick, and prepared by vibratome) through the ventral lateral thalamus of *Cdx2^{Flpo}-tdTomato* mice (left, $n = 3$) and *Tac1^{Cdx2}-tdTomato* mice (right, $n = 3$). *Cdx2^{Flpo}-tdTomato* mice were generated by crossing *Cdx2^{Flpo}* mice with Flpo-dependent *Rosa26$^{FSF\text{-}tdTomato}$* reporter mice. Among 25-μm-thick sections of VPL from *Tac1^{Cdx2}-tdTomato* mice, only 12% (3 out of 26) showed sparse tdTomato signals, whereas 100% (26 out of 26) of sections from *Cdx2^{Flpo}-tdTomato* mice showed robust tdTomato signals (χ^2 test, $\chi^2_{0.95,(1)} = 41.241$, $P < 0.001$). It should be noted that owing to the restriction of CDX2^{Flpo} to the spinal cord, no tdTomato signals were detected in VPM of *Cdx2^{Flpo}-tdTomato* mice, because VPM are innervated by neurons located in the trigeminal nuclei or dorsal column nuclei that were not labelled by CDX2^{Flpo}. **g**, **h**, Representative coronal sections (100-μm thick) through the thalamus of P30 *Cdx2^{Flpo}-tdTomato* mice ($n = 2$) at the level of bregma −1.70 mm, showing whole spinal ascending fibres in the medial thalamic complex. **i**, Representative coronal sections (25 μm) of PBN from P30 *Cdx2^{Flpo}-tdTomato* mice ($n = 2$), showing tdTomato (red) and CGRP immunostaining (green). CDX2^{Flpo}–tdTomato^{+} fibres send collateral terminals to CGRP^{+} PBel as indicated by the arrow, besides projections to PBel and PBdvl. Scale bars, 50 μm (**a–c**), 100 μm (**e–i**).

© 2019 Springer Nature Limited. All rights reserved.

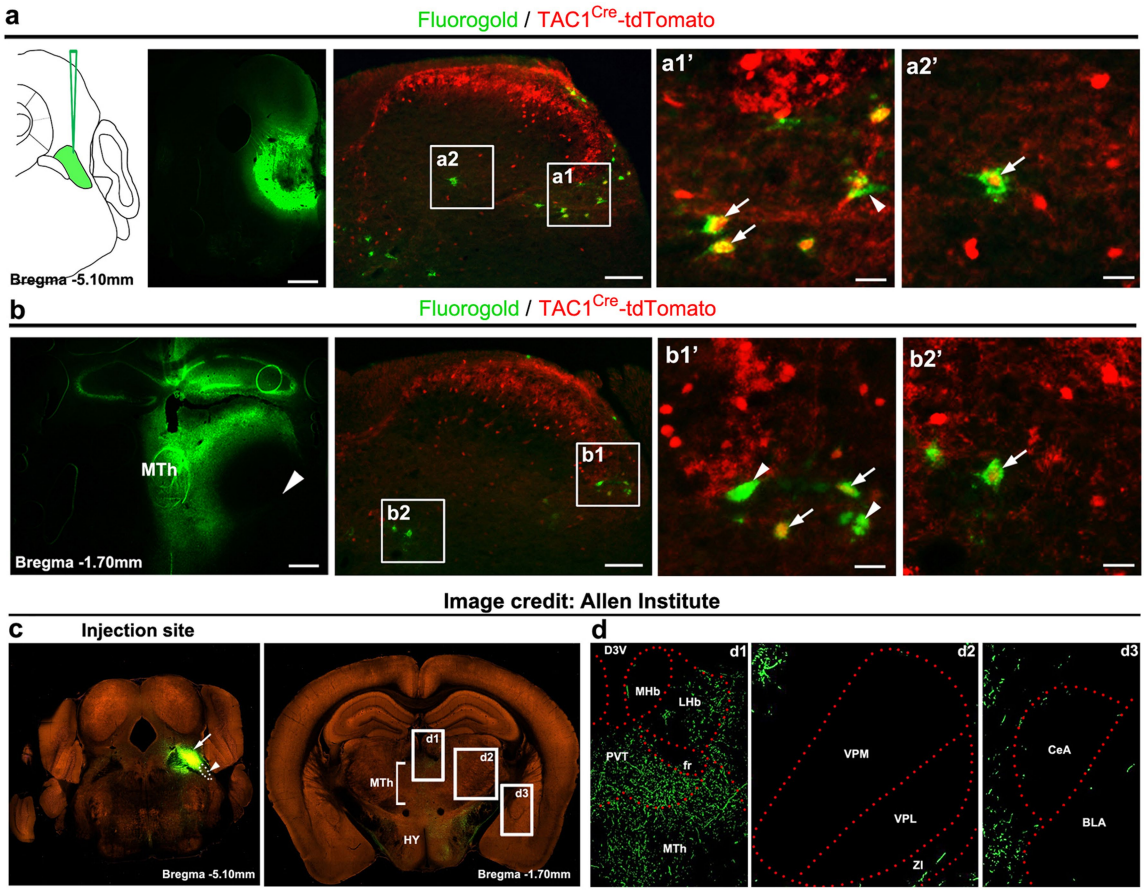


Extended Data Fig. 2 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

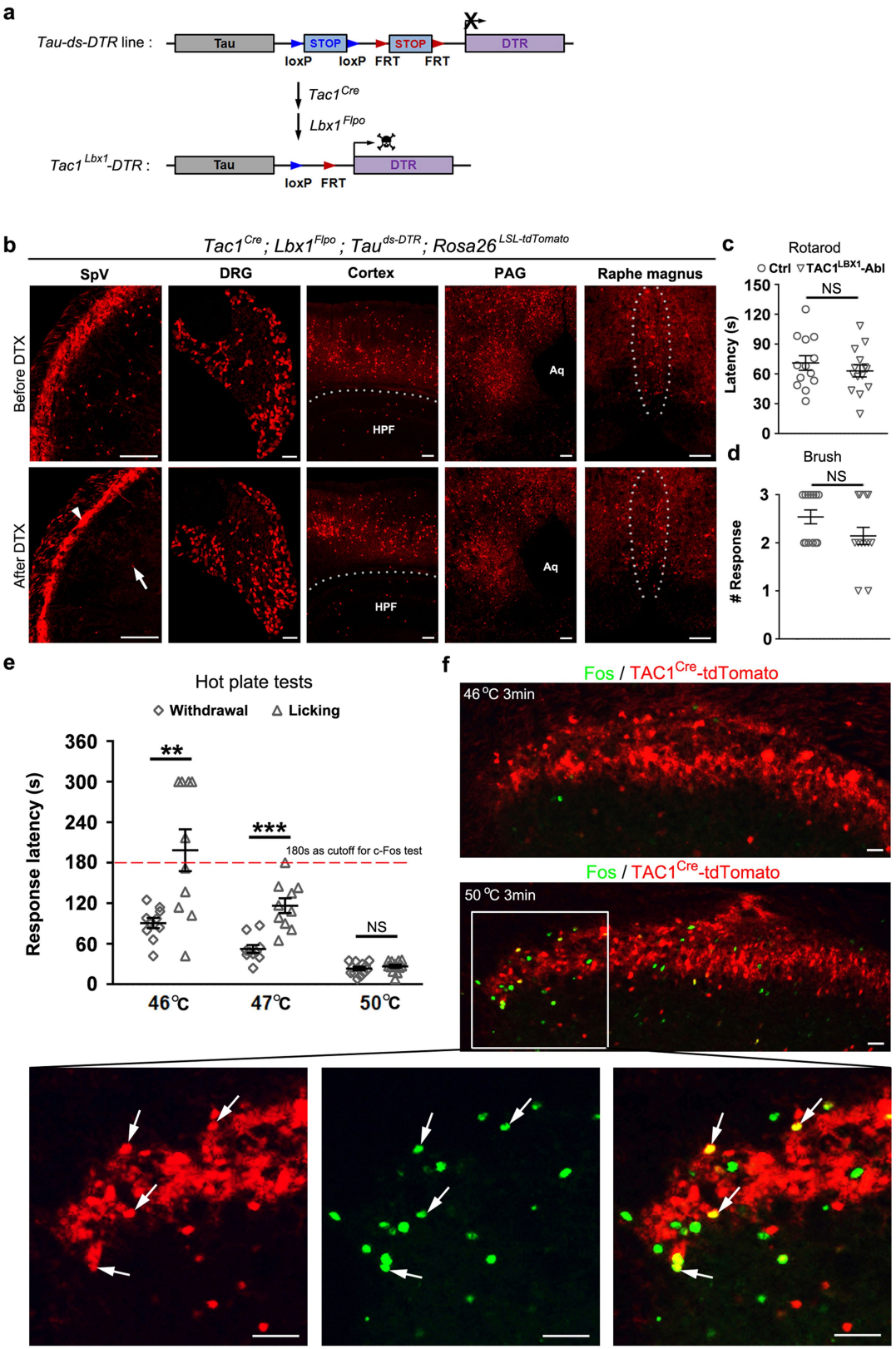
Extended Data Fig. 2 | Functional connections of *Tac1*^{cre}-marked neurons to neurons in PBsl. a, Representative images showing the distribution of presynaptic reporter (the synaptophysin–EGFP fusion protein)[12] in the dorsal portion of lateral PBN—including PBsl (**a**, middle, green)—but not in the more ventral PBel, following intraspinal injection of the AAV-Syn1-DIO-tdTomato-T2A-SynEGFP virus at the lumbar level of adult *Tac1*^{cre} mice ($n = 2$). The TAC1^{Cre+} axons are visualized by tdTomato signals (red). Arrowhead indicates potential axons that pass through ventral lateral PBN without making synapses. **b**, Intersectional genetic strategy for driving the expression of the calcium translocating channelrhodopsin (CatCh, an L132C mutant channelrhodopsin with enhanced Ca^{2+} permeability and fused with GFP)[43] in spinal TAC1^{CDX2} neurons defined by co-expression of TAC1^{Cre} and CDX2^{Flpo}. This was achieved by crossing the intersectional CatCh mice (*Ai80*) with *Tac1*^{cre} and *Cdx2*^{Flpo}, with the resulting triple heterozygous mice referred to as *Tac1*^{Cdx2}*-CatCh* mice. Triple heterozygous *Tac1*^{Cdx2}*-GFP* mice—generated by crossing *Tac1*^{cre} and *Cdx2*^{Flpo} mice with intersectional GFP reporter mice *RC:: FrePe*—were used as control. **c**, The ascending TAC1^{CDX2}–CatCh–GFP^{+} terminals (observed from 3 mice) were detected in the medial thalamic region (left), including the PVT and MTh. Right, GFP signals in PBsl. The fluorescent signal of CatCh–GFP fusion protein detected by immunostaining is not as robust as that shown by the direct visualization of tdTomato signals observed in *Tac1*^{Cdx2}*-tdTomato* mice, in Fig. 1. Scale bars, 100 μm. **d**, Top, schematic (left) of optogenetic activation of TAC1^{CDX2}–CatCh^{+} terminals in PBN via 473-nm blue light, and recording sites for neurons in PBsl or PBel (middle and right, respectively) of the same brain slices. Bottom, voltage clamp (V-clamp) was used to record the evoked excitatory postsynaptic currents (EPSC), with holding membrane potential (HP) at −70 mV. Current clamp was used to record action potential (AP) firing. RP, resting membrane potential. Representative recording traces show that neurons in PBsl but not in PBel responded to the blue-light stimulation (0.2 Hz, 20 ms, numbers of neurons with responses: PBsl, 4 out of 15; PBel, 0 out of 15; χ^2 test, $\chi^2_{0.95,(1)} = 4.615$, $P = 0.032$; *Tac1*^{Cdx2}*-CatCh* mice, $n = 2$).

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 3 | Retrograde labelling of spinal *Tac1*^{cre}-marked projection neurons from parabrachial and medial thalamic nuclei, and anterograde tracing from the dorsal part of lateral parabrachial nuclei. a, Fluorogold retrograde labelling from PBN of *Tac1*^{cre}-*tdTomato* mice ($n = 3$). Left, injection site. Middle and right, representative transverse section of the dorsal horn showing fluorogold$^+$ retrograde-labelled cells (green) and tdTomato$^+$ TAC1-lineage neurons (red). Arrows indicate colocalization in the lateral spinal nucleus (a1) and deep laminae (a2). Arrowhead indicates a tdTomato-negative retrograde labelled neuron, showing that TAC1^{Cre}-marked neurons represent a subset of spinoparabrachial projection neurons ($n = 3$, 27.2 ± 0.7%). **b**, Fluorogold retrograde labelling from the medial thalamic nuclei ($n = 3$ mice). Left, injection site. Large arrowhead indicates that the fluorogold injection did not leak to the lateral VPM–VPL complex. Middle and right, a representative transverse section of the dorsal horn, showing fluorogold$^+$ retrograde-labelled cells (green) and tdTomato$^+$ *Tac1*-lineage neurons (red). Arrows indicate colocalization in the lateral spinal nucleus (b1) and deep laminae (b2). Small arrowheads indicate tdTomato-negative retrograde-labelled neurons, indicating that TAC1^{Cre}-marked neurons again represent a subset of spinothalamic projection neurons ($n = 3$ mice, 16.3 ± 3.8%). **c**, Anterograde tracing from dorsal lateral PBN (including PBsl and PBdvl). Image credit, Allen Institute. Left, the coronal plane (bregma −5.10 mm) showing the injection site in parabrachial nuclei. Arrow indicates tracer injection confined to PBsl plus PBdvl. The white-dotted circle (arrowhead) indicates the PBel that contains little or no injected tracer. Right, the projection of neurons from PBsl and PBdvl to thalamic and hypothalamic regions (bregma −1.70 mm). Boxes d1, d2 and d3 highlight projections or lack of projections to medial thalamic nuclei, ventral lateral thalamic nuclei and amygdaloid nuclei shown in **d**, respectively. HY, hypothalamic nuclei. **d**, Dense innervations were observed in medial thalamic nucleus, lateral habenular nuclei and the paraventricular nucleus of the thalamus (d1). No innervations were observed in VPM or VPL (d2), or the central (CeA) and basal lateral (BLA) parts of amygdala (d3). The lack of innervations to CeA, which is innervated by CGRP$^+$ neurons in PBel[8], provided a further indication that the tracer injection to the PBsl–PBdvl region did not diffuse to the PBel region. The full set of tracing images is available at the Allen Mouse Brain Connectivity Atlas[12] (http://connectivity.brain-map.org/projection/experiment/siv/127469566?imageId=127469776&imageType=TWO_PHOTON,SEGMENTATION&initImage=TWO_PHOTON&x=18728&y=17591&z=3; injection site picture, modified from image 104 of 140; thalamic projection picture, modified from image 71 of 140).

© 2019 Springer Nature Limited. All rights reserved.

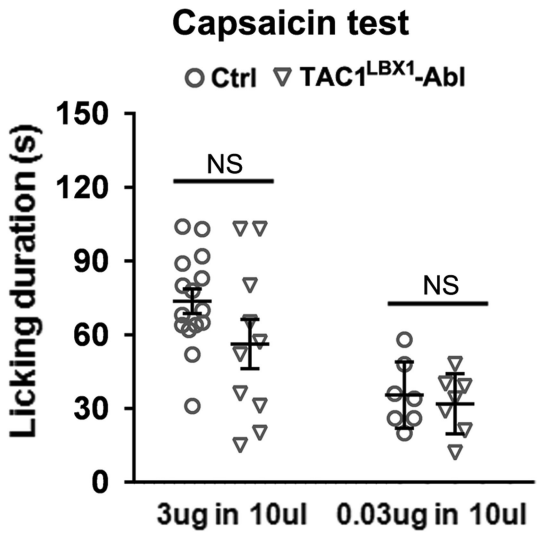


Extended Data Fig. 4 | See next page for caption

© 2019 Springer Nature Limited. All rights reserved.

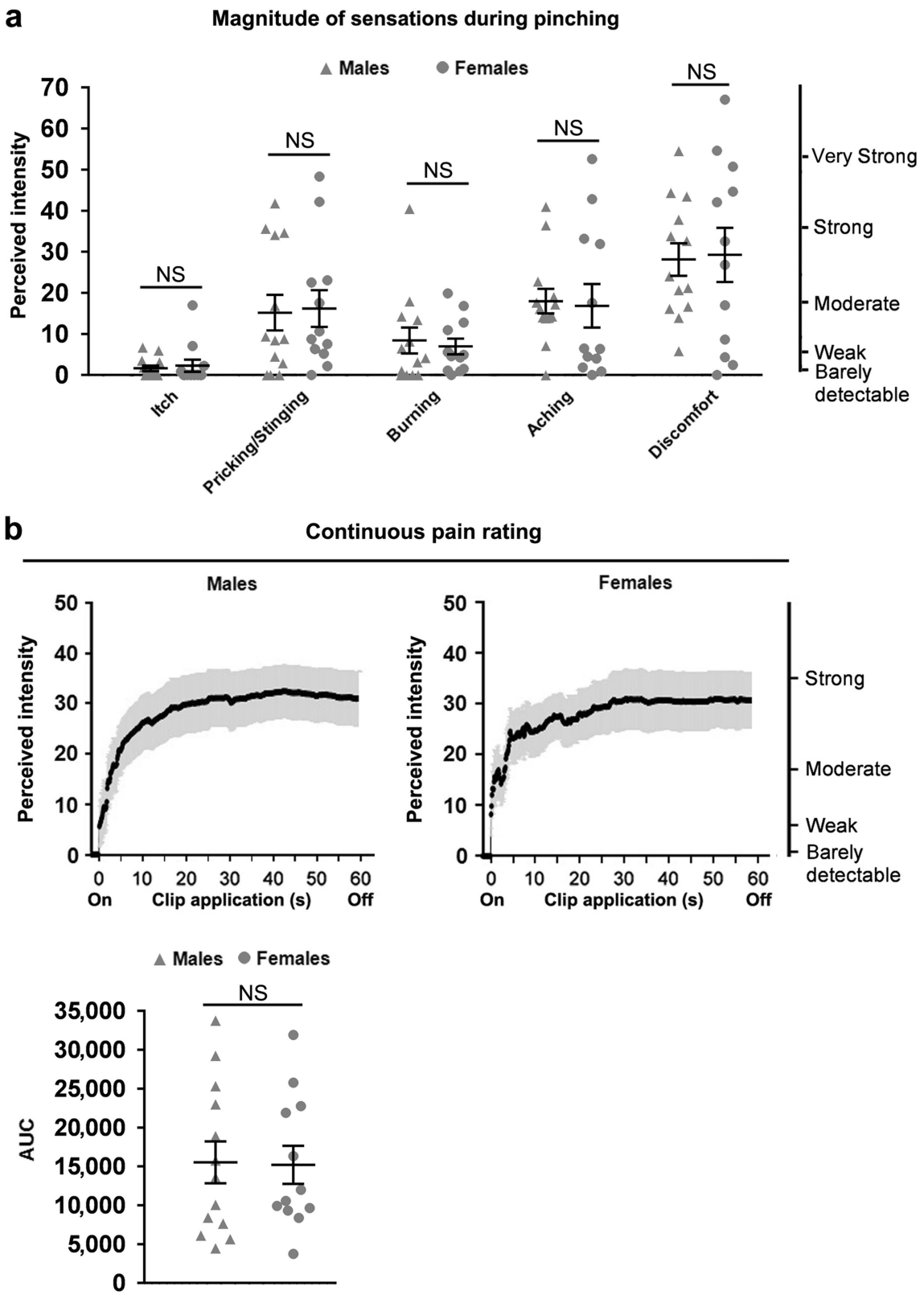
Extended Data Fig. 4 | Additional anatomical and behavioural characterizations of mice in which TAC1^{LBX1} neurons were ablated, as well as temporal segregation of withdrawal versus licking responses evoked by noxious heat and their correlation with Fos induction in TAC1^{Cre}-marked neurons. a, Intersectional genetic strategy for driving DTR expression selectively in dorsal spinal cord TAC1^{LBX1} neurons defined by co-expression of TAC1^{Cre} and LBX1^{Flpo}. DTR is driven from the pan-neural promoter *Tau*, and its expression requires removal of two STOP cassettes by Cre and Flpo DNA recombinases. LBX1^{Flpo} expression is confined to the dorsal hindbrain and dorsal spinal cord within the nervous system[6,13]. **b**, Representative images showing a marked loss of tdTomato^{+} cells in the hindbrain spinal trigeminal nucleus (SpV) after diphtheria toxin injections ($n = 3$ mice). Arrow in SpV indicates one of few remaining cells; arrowhead indicates processes derived from TAC1^{Cre}–tdTomato^{+} trigeminal primary afferents that were preserved. TAC1^{Cre}–tdTomato^{+} neurons are preserved in dorsal root ganglia (DRG) and trigeminal ganglia (not shown), as well as in various brain regions such as the cortex, hippocampus formation (HPF), periaqueductal grey nuclei (PAG) or raphe magnus. Aq, aqueduct. **c**, No detected difference in falling latencies from the rotarod between control littermates and TAC1^{LBX1}-Abl mice (control, $n = 13$ mice; TAC1^{LBX1}-Abl, $n = 14$ mice; two-sided t-test, $P = 0.403$). **d**, No detected difference in response rates to gentle hindpaw brushing (out of three tries for each mouse) between control and TAC1^{LBX1}-Abl groups (control, $n = 13$ mice; TAC1^{LBX1}-Abl, $n = 14$ mice; two-sided Mann–Whitney rank-sum test, $P = 0.121$). **e**, Wild-type mice showed distinct latencies of lifting or flinching versus licking in response to hot plate stimulation set at 46–47 °C, but no difference at 50 °C (46 °C, $n = 10$ mice, two-sided paired t-test, $P = 0.006$; 47 °C, $n = 10$ mice, two-sided paired t-test, $P < 0.001$; 50 °C, $n = 12$ mice, two-sided paired t-test, $P = 0.379$). In the 46-°C hot plate test, licking responses were rarely observed within the first 3 min. **f**, Representative immunostaining of Fos in superficial dorsal horn of *Tac1^{cre}-tdTomato* mice 2 h after 3-min exposure to the 46-°C or 50-°C hot plate ($n = 3$ mice for each condition). Only 50 °C could induce robust Fos expression. Bottom panels represent the boxed area shown above. Arrows indicate colocalization of Fos (green) with TAC1^{Cre}-marked tdTomato^{+} cells (red). Scale bars, 100 μm (**b**), 25 μm (**f**). Data are presented as mean ± s.e.m. (**c**, **e**) or median ± quartile (**d**).

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Mice in which TAC1^{LBX1} neurons were ablated still produced licking responses to intraplantar capsaicin injection. No difference in licking evoked by 10 μl of a solution that contained 3 μg or 0.03 μg capsaicin (3-μg test, control, $n = 15$ mice; TAC1^{LBX1}-Abl, $n = 10$ mice, two-sided t-test, $t_{(23)} = 1.714$, $P = 0.143$; 0.03-μg test, control, $n = 7$ mice; TAC1^{LBX1}-Abl, $n = 7$ mice, two-sided t-test, $t_{(12)} = 0.519$, $P = 0.613$), suggesting the existence of pain pathways that are independent of TAC1^{LBX1} neurons. This preservation of capsaicin-evoked licking is markedly different from a complete loss of licking evoked by mustard oil and other noxious stimuli (Fig. 3). Licking responses evoked by mustard oil at a low concentration ($\leq$0.75%) are dependent on TRPA1[44], and TRPA1 is expressed in a subset of TRPV1^{+} neurons[45]. As such, neurons that are responsive to mustard oil represent only a subset of capsaicin-responsive neurons[46,47]. In other words, there are neurons that are sensitive to capsaicin and insensitive to mustard oil that could—in principle—mediate licking that is independent of TAC1^{LBX1} neurons, and which is evoked by capsaicin. Data shown as mean ± s.e.m.

© 2019 Springer Nature Limited. All rights reserved.

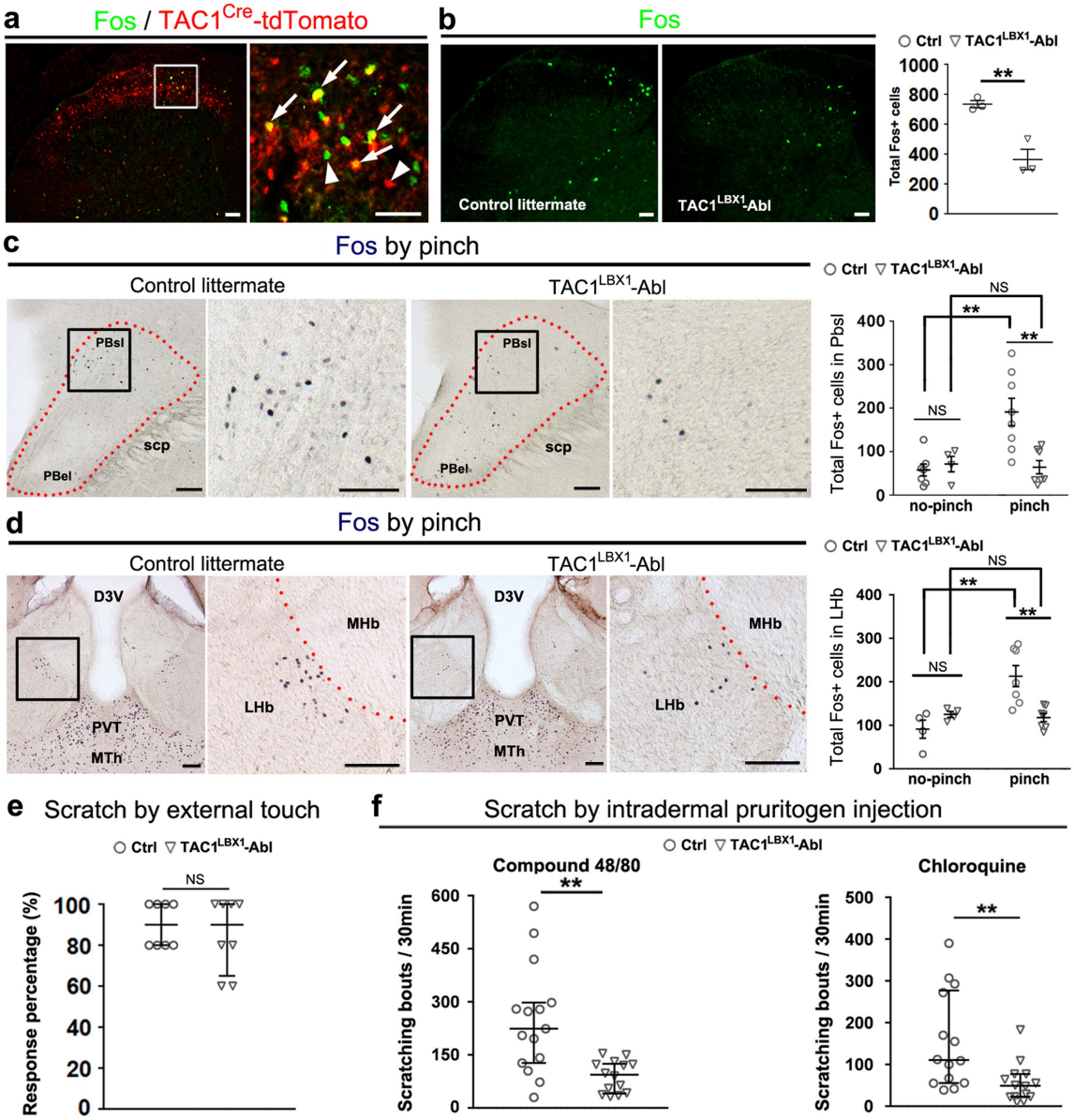


Extended Data Fig. 6 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

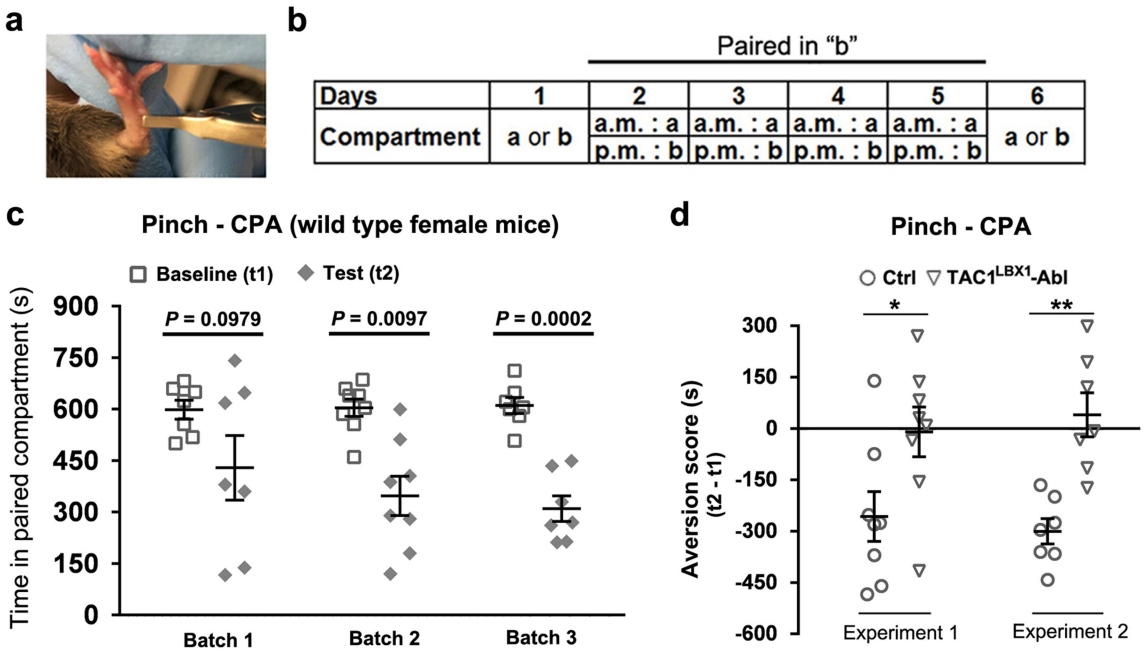
Extended Data Fig. 6 | Skin-pinch-evoked sustained pain in humans. During the application of the alligator clip, both female and male subjects were instructed to rate continuously the perceived intensity of pain, regardless of its quality. After the clip was removed, each subject was asked to rate—in similar fashion—the maximal perceived intensity of each of four aversive qualities of cutaneous sensation associated with the pain they had just experienced. The four sensory qualities were itch, pricking or stinging, burning, and aching. Then, subjects were asked to rate the discomfort associated with this maximal sensation. The common scale at the right side indicates the intensity of each sensation (see Methods for detail). **a**, No differences between male ($n = 13$) and female ($n = 12$) human subjects in rating the magnitude of the indicated sensory qualities (two-sided Mann–Whitney rank-sum test; itch, $U = 75.0$, $P = 0.874$; pricking or stinging, $U = 69.5$, $P = 0.663$; burning, $U = 71.0$, $P = 0.723$; and aching, $U = 62.5$, $P = 0.414$; two-sided t-test, discomfort, $t_{(23)} = -0.150$, $P = 0.882$; data shown as mean ± s.e.m.). **b**, No differences in the continuous pain rating between males ($n = 13$) and females ($n = 12$) (top panels, continuous pain rating at different time points during the one-minute pinch period were subjected to two-way ANOVA analyses with repeated measures, and no significant difference was detected between genders, $F_{(1,23)} = 0.008$, $P = 0.929$; bottom, the areas under the entire curve (AUC) did not show a difference between genders, two-sided t-test, $t_{(23)} = 0.089$, $P = 0.929$, data shown as mean ± s.e.m.). This lack of any detectable gender differences with the current sample sizes is consistent with previous studies that show that gender differences for experimentally evoked pain are not easy to detect in humans[48,49].

© 2019 Springer Nature Limited. All rights reserved.



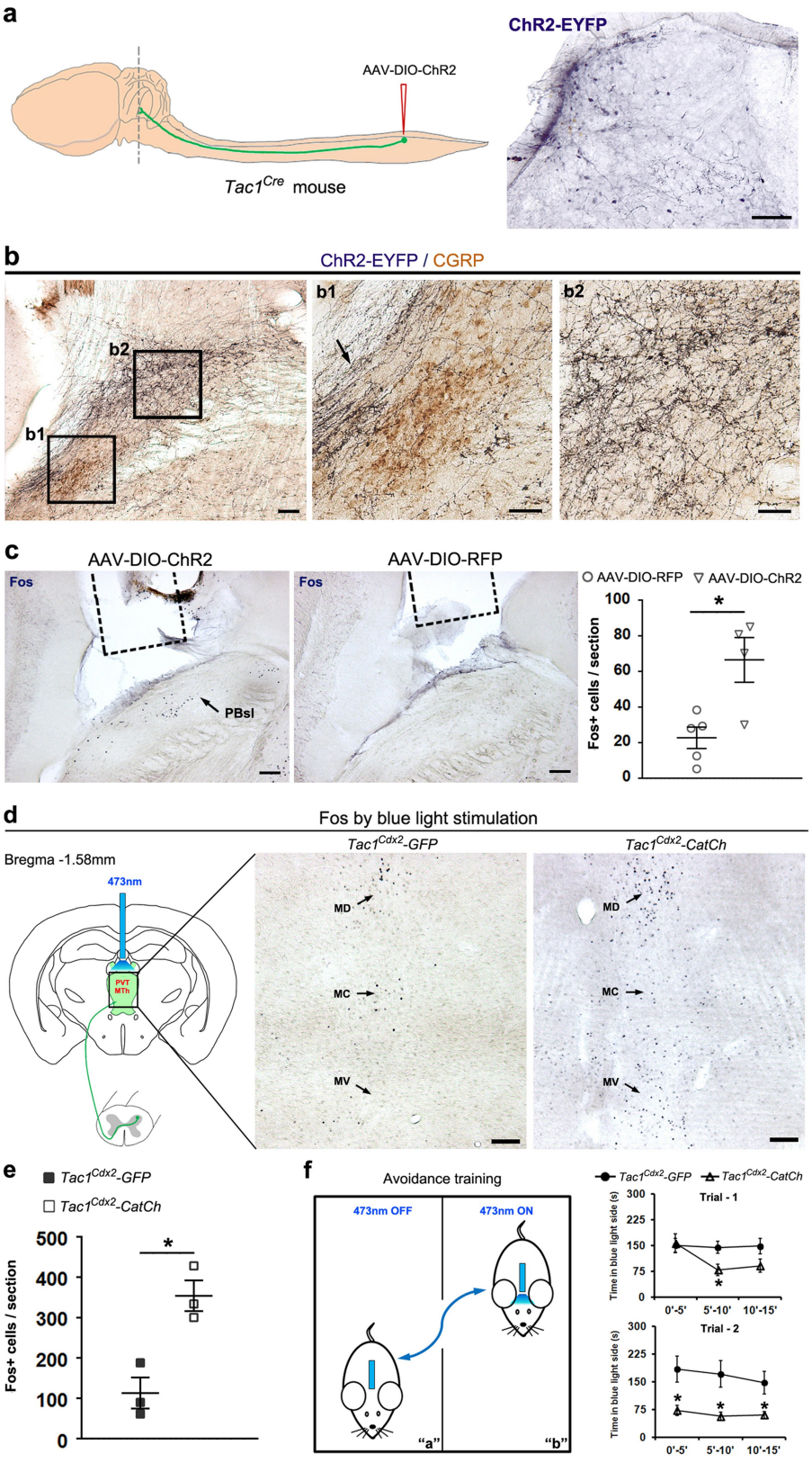
Extended Data Fig. 7 | Loss of pinch-induced Fos expression in the dorsal horn, PBsl and LHb, and attenuated pruritogen-induced scratching in mice in which TAC1^{LBX1} neurons were ablated. **a**, Representative lumbar spinal cord sections of P60 *Tac1^{cre}-tdTomato* mice ($n = 4$) after hindpaw pinch stimulation. We found that 41.7 ± 8.0% neurons with pinch-induced Fos co-expressed tdTomato. Arrows indicate co-localization and arrowheads indicate singular expression. **b**, Reduced Fos in lumbar dorsal horn of TAC1^{LBX1}-Abl mice ($n = 3$ mice for each group, two-sided *t*-test, $P = 0.007$). **c**, Representative images showing pinch-induced Fos on coronal sections through the lateral PBN. Note that in wild-type littermates, pinch-induced Fos was enriched in PBsl, and only rarely in PBel. Right, quantification of Fos$^+$ cells between bregma −5.24 and −4.96 mm, with and without pinching (no-pinch group, control littermates, $n = 7$; TAC1^{LBX1}-Abl, $n = 4$ mice; pinch group, control littermates, $n = 8$; TAC1^{LBX1}-Abl, $n = 7$ mice). Two-way ANOVA indicates significant interactions between genotypes and pinch stimulation ($F_{(1,22)} = 8.555$, $P = 0.008$); post hoc comparison (Holm–Sidak method) shows comparable basal levels of Fos expression (no-pinch groups, $P = 0.72$), an increase in control littermates within PBsl ($P = 0.004$) and the loss of this increase in TAC1^{LBX1}-Abl mice ($P = 0.006$). **d**, Representative coronal sections through the dorsal midline thalamic complex, showing bilateral Fos induction by pinch, which is consistent with previous electrophysiological studies[50,51]. Right, counting of pinch-induced Fos$^+$ cells in a region of LHb that is adjacent to MHb from bregma −1.46 to −2.06 mm (no-pinch groups, control littermates, $n = 4$; TAC1^{LBX1}-Abl, $n = 4$ mice; pinch groups, control littermates, $n = 7$; TAC1^{LBX1}-Abl, $n = 7$ mice). Two-way ANOVA indicates significant interactions between genotypes and pinch stimulation ($F_{(1,18)} = 11.08$, $P = 0.004$); post hoc comparison (Holm–Sidak method) shows comparable basal-level Fos expression (no-pinch groups, $P = 0.289$), significant increase in LHb of control littermates ($P = 0.008$) and loss of this increase in TAC1^{LBX1}-Abl mice ($P = 0.003$). Owing to high background expression of Fos in the PVT and MTh, we cannot determine pinch-evoked neuronal activation in these nuclei. **e**, No difference in scratching response rates evoked by light von-Frey-filament stimulation (control, $n = 8$ mice; TAC1^{LBX1}-Abl, $n = 8$ mice; two-sided Mann–Whitney rank-sum test, $P = 0.721$). **f**, Reduced scratching bouts induced by intradermal pruritogen injection (compound 48/80 (a polymer produced by the condensation of *N*-methyl-*p*-methoxyphenethylamine with formaldehyde) test, control, $n = 15$ mice; TAC1^{LBX1}-Abl, $n = 14$ mice; two-sided Mann–Whitney rank-sum test, $P = 0.002$; chloroquine test, control, $n = 14$ mice; TAC1^{LBX1}-Abl, $n = 14$ mice; two-sided Mann–Whitney rank-sum test, $P = 0.005$). Ctrl, control littermates. NS, $P > 0.05$. Data shown as mean ± s.e.m. (**b**–**d**) or mean ± quartile (**e**, **f**). Scale bars, 50 μm (**a**, **b**), 100 μm (**c**, **d**).

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Loss of pinch-induced CPA in female mice in which TAC1^{LBX1} neurons were ablated. a, A pinched mouse hindpaw. The alligator clip was applied to the ventral skin surface between the footpad and the heel. **b**, The experimental programme for pinch-evoked CPA test (for details, see Methods). **c**, Hindpaw skin pinch, but not sham handling (grabbing without pinching, data not shown), induced CPA in wild-type females. CPA is measured by the change in the amount of time mice stay in the paired chamber before (baseline, t_1) and after (test, t_2) pinch-evoked conditioning. In three independent batches of wild-type control littermates, a two-sided *t*-test showed that the second and third batches—but not the first batch—displayed significant reduction of t_2 in comparison with t_1 (two-sided *t*-test, batch 1, $n = 7$ mice, $P = 0.0979$; batch 2, $n = 8$ mice, $P = 0.0097$; batch 3, $n = 7$ mice, $P = 0.0002$). Two-way ANOVA analyses of these three batches indicated that pinch-evoked avoidance of the paired chamber ($F_{(1,19)} = 36.514$, $P < 0.001$), without showing batch effects ($F_{(2,19)} = 0.547$, $P = 0.587$) and interactions ($F_{(2,19)} = 0.885$, $P = 0.429$). This suggests that pinching can induce CPA in wild-type females. **d**, Female mice in which TAC1^{LBX1} neurons were ablated showed a loss of pinch-induced CPA (control littermates and TAC1^{LBX1}-Abl mice, $n = 8$ for experiment 1, *t*-test, $^*P = 0.031$; $n = 7$ for experiment 2, *t*-test, $^{**}P = 0.001$); this was also true for male mice in which TAC1^{LBX1} neurons were ablated (Fig. 3).

© 2019 Springer Nature Limited. All rights reserved.

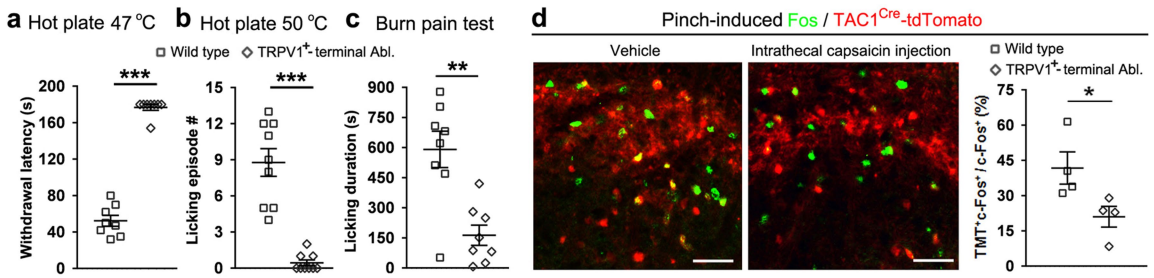


Extended Data Fig. 9 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Fig. 9 | Optogenetic activation of terminals derived from TAC1^{Cre}-marked neurons around PBN or the medial thalamus. **a**–**c**, Viral infection in lumbar spinal cord TAC1^{Cre+} neurons and subsequent optogenetic activation of the central terminals in PBN. **a**, The AAV-DIO-ChR2 virus, which drives the expression of the fusion ChR2–EYFP protein in a Cre-dependent manner, was injected into the lumbar dorsal horn of *Tac1^{cre}* mouse, and the ascending ChR2–EYFP^{+} terminals around PBN (dashed line) were visualized by a GFP antibody. These mice are referred to as TAC1^{Cre}–ChR2 mice. Right, immunostaining shows expression of the ChR2–EYFP fusion protein in the lumbar spinal cord. **b**, Representative images showing double-colour immunostaining, which reveals the ascending projections to the PBN at bregma −5.02 mm. TAC1^{Cre}–ChR2–EYFP^{+} terminals (dark blue) were co-stained with CGRP (brown). Note that the TAC1^{Cre}–ChR2–EYFP^{+} fibres pass through a region (b1, arrow) lateral to CGRP^{+} PBel (b1, brown), and terminated densely in PBsl (b2). **c**, Representative images showing Fos expression induced by blue-light stimulation, in PBsl of TAC1^{Cre}–ChR2 mice; there are far fewer Fos^{+} neurons after blue-light stimulation in control TAC1^{Cre}–RFP mice, in which viral injection drove the expression of RFP but not ChR2 (ChR2 mice, $n = 4$; RFP-control mice, $n = 5$; two-sided t-test, $P = 0.012$). Dashed lines indicate the location of the implanted optic fibre in the region above the PBN. **d**–**f**, Optogenetic experiments for spinal TAC1 neurons projected to medial thalamic nuclei. **d**, **e**, Generation of the intersectional *Tac1^{Cdx2}-CatCh^{+}* mice is described in Extended Data Fig. 2b. The optic fibre was implanted above the medial thalamic complex (left). Neuronal activation by blue-light stimulation is indicated by the increase of Fos^{+} cells in *Tac1^{Cdx2}-CatCh* mice in comparison with *Tac1^{Cdx2}-GFP* mice, as shown by representative images and quantitative analyses (*Tac1^{Cdx2}-CatCh*, $n = 3$ mice; *Tac1^{Cdx2}-GFP*, $n = 3$ mice; two-sided t-test, $P = 0.011$). **f**, The *Tac1^{Cdx2}-CatCh* mice showed a progressive avoidance of the blue-light-paired chamber during two fifteen-minute training trials conducted on two consecutive days (see Methods for details). A two-way ANOVA plus post hoc Bonferroni's t-test showed a progressive avoidance of the paired chamber (*Tac1^{Cdx2}-CatCh* mice, $n = 10$; *Tac1^{Cdx2}-GFP* control mice, $n = 11$; trial 1, significant interaction, $F_{(2,38)} = 5.067$, $P = 0.011$; trial 2, significant genotype effect, $F_{(1,19)} = 6.825$, $P = 0.017$, no interaction, $F_{(2,38)} = 0.73$, $P = 0.489$).

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Ablation of TRPV1$^+$ central terminals led to impaired responses to noxious heat or skin burn injury, as well as a reduction of pinch-induced Fos expression in TAC1^{Cre}-marked neurons in the dorsal horn. a, Mice in which TRPV1$^+$ central terminals were ablated showed a marked increase in the withdrawal latency evoked by the 47-°C hot plate stimulation, with cut off time set at 3 min (vehicle injection versus intrathecal capsaicin injection groups, $n = 8$ mice for each group, two-sided Mann–Whitney rank-sum test, $^{***}P < 0.001$). This is stark contrast to subtle and insignificant changes seen in mice in which TAC1^{LBX1} neurons were ablated (Fig. 2e). **b**, Loss of licking behaviour evoked by the 50-°C hot plate stimulation (vehicle injection versus intrathecal capsaicin injection groups, $n = 9$ mice for each group; licking episodes within one minute, two-sided Mann–Whitney rank-sum test, $^{***}P < 0.001$). **c**, Mice in which TRPV1$^+$ central terminals were ablated also displayed a marked reduction in the licking evoked by hindpaw burn injury ($n = 8$ mice for each group, licking duration within 30 min of skin burn injury, two-sided t-test, $P = 0.001$). **d**, Representative immunostaining images and quantitative analyses showing pinch-evoked Fos expression (green) in the dorsal horn of *Tac1^{cre}-tdTomato* mice, with or without chemical ablation of TRPV1$^+$ central terminals ($n = 4$ mice for each group). Note that there is a reduction of pinch-induced Fos expression in TAC1^{Cre}-marked tdTomato$^+$ cells after ablation of TRPV1$^+$ central terminals (two-sided t-test, $P = 0.044$). TRPV1$^+$ nociceptors are necessary for both reflexes[20] (**a**) and licking (**b**, **c**) evoked by noxious heat. Earlier studies have shown that the dorsal root ganglia neurons with highest TRPV1 expression (TRPV1^{highest}), which represent about 10% of TRPV1$^+$ nociceptors[52], respond to moderately hot stimulation[53]. Similar to MRGPRD$^+$ nociceptors, these TRPV1^{highest} neurons innervate exclusively the skin epidermis[53], and their development is dependent on the same transcription factor, RUNX1[52,54]. We therefore speculate that these TRPV1^{highest} neurons may be involved with the first-line reflexes evoked by noxious heat. This raises the possibility that there are different subsets of TRPV1$^+$ nociceptors that are associated with reflexes versus sustained pain evoked by noxious heat: future experiments are needed to test this hypothesis.

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0785-8

A male-expressed rice embryogenic trigger redirected for asexual propagation through seeds

Imtiyaz Khanday[1,2], Debra Skinner[1], Bing Yang[3], Raphael Mercier[4] & Venkatesan Sundaresan[1,2,5]*

The molecular pathways that trigger the initiation of embryogenesis after fertilization in flowering plants, and prevent its occurrence without fertilization, are not well understood[1]. Here we show in rice (*Oryza sativa*) that BABY BOOM1 (BBM1), a member of the AP2 family[2] of transcription factors that is expressed in sperm cells, has a key role in this process. Ectopic expression of *BBM1* in the egg cell is sufficient for parthenogenesis, which indicates that a single wild-type gene can bypass the fertilization checkpoint in the female gamete. Zygotic expression of *BBM1* is initially specific to the male allele but is subsequently biparental, and this is consistent with its observed auto-activation. Triple knockout of the genes *BBM1*, *BBM2* and *BBM3* causes embryo arrest and abortion, which are fully rescued by male-transmitted *BBM1*. These findings suggest that the requirement for fertilization in embryogenesis is mediated by male-genome transmission of pluripotency factors. When genome editing to substitute mitosis for meiosis (*MiMe*)[3,4] is combined with the expression of *BBM1* in the egg cell, clonal progeny can be obtained that retain genome-wide parental heterozygosity. The synthetic asexual-propagation trait is heritable through multiple generations of clones. Hybrid crops provide increased yields that cannot be maintained by their progeny owing to genetic segregation. This work establishes the feasibility of asexual reproduction in crops, and could enable the maintenance of hybrids clonally through seed propagation[5,6].

Understanding the molecular pathway that underlies the initiation of embryogenesis by a fertilized egg cell is a major unresolved problem in plant development[1]. In animals, the initiation of embryogenesis depends upon defined maternal factors[7]. In plants, two contrasting models have been proposed: one suggests that the two parental genomes contribute equally[8], whereas the other considers that the maternal genome has the primary role in early embryogenesis[9,10]. The identity and parental origin of the specific factors in plants that trigger zygotic development are as yet undetermined. We have previously used rice to elucidate transcriptome dynamics during the zygotic transition[11] and found that *BABY BOOM* (*BBM*)-like transcription factors of the APETALA 2/ETHYLENE RESPONSE FACTOR (AP2/ERF) superfamily[12] are expressed in zygotes after fertilization, which suggests a potential role in the initiation of embryogenesis (Extended Data Table 1a). *BBM* genes from *Arabidopsis thaliana* and *Brassica napus* can ectopically induce somatic embryos[13]; however, a role for these genes in the initiation of zygotic embryos has not been established[2]. We first determined that ectopic expression of *BBM1*—a *BBM*-like gene expressed in rice zygotes—also resulted in somatic embryos, both by examining their morphology and by using embryo marker genes (Extended Data Fig. 1a–d). Because *BBM1* expression increases with the age of the zygote[11] (Extended Data Table 1a), we investigated whether its expression is autoregulated, by inducing a constitutive BBM1–glucocorticoid receptor (GR) fusion in somatic tissues using dexamethasone (DEX) (Extended Data Fig. 1e). Quantitative PCR after reverse transcription (RT–qPCR), using allele-specific primers, showed that the expression of endogenous *BBM1*—but not the *BBM1-GR* fusion transgene—was highly induced after 24 h of DEX treatment (Extended Data Fig. 1f–h). This expression was maintained in the presence of the protein-biosynthesis inhibitor cycloheximide (CYC), indicating that BBM1 auto-activation is likely to be direct (Extended Data Fig. 1h). Auto-activation might be a conserved feature of *BBM* genes, because *B. napus* BABY BOOM can activate the expression of *Arabidopsis BBM*[14].

Our previous study of hybrid zygote transcriptomes[11] indicated that, although most zygotic transcripts were from the female genome, a few de novo transcription factors—including *BBM1*—had male-derived transcripts. We used RT–PCR amplification across single nucleotide polymorphisms (SNPs) in *BBM1* to confirm that, at 2.5 h after pollination (HAP) (corresponding to karyogamy), only the male *BBM1* allele is expressed in reciprocal crosses of *indica* and *japonica* cultivars[11] (Extended Data Fig. 2a). These results were confirmed in isogenic zygotes in the *japonica* Kitaake cultivar. We reciprocally crossed wild-type plants to transgenic plants that carried a translational fusion of the *BBM1* genomic locus to GFP (*BBM1-GFP*) (Extended Data Fig. 2b). Zygotes at 2.5 HAP displayed GFP expression only if the *BBM1-GFP* transgene was transmitted from the male parent (Fig. 1a). Consistent with this observation, in *BBM1-GFP* selfed progeny, GFP was detected in only about half of the zygotes, instead of the three-quarters ratio that would be expected if there is no parent-of-origin bias (Fig. 1a). Subsequently, GFP expression can be detected from the female allele in 6.5 HAP zygotes, corresponding to mid-to-late G2 phase (Extended Data Fig. 2c, d). Because BBM1 is capable of auto-activation of its own promoter (Extended Data Fig. 1h), the late expression of *BBM1* from the female allele might result from earlier expression of *BBM1* from the male allele. Other redundantly acting *BBM* genes might also contribute to this delayed activation (see below). *BBM1* expression continues through the later stages of embryo development (Extended Data Fig. 2e). In gametes, *BBM1* RNA can be detected by RT–PCR in sperm cells but not in egg cells (Extended Data Fig. 2f), which is consistent with RNA sequencing data[15] (Extended Data Table 1a). Furthermore, the BBM1–GFP fusion protein was expressed in sperm cells, which suggests that both transcription and translation of *BBM1* can occur in male gametes before fertilization (Extended Data Fig. 2g).

The expression of *BBM1* specifically from the male genome after fertilization, together with its capability to induce somatic embryogenesis, suggested that BBM1 could be a trigger of embryo development in the zygote (Extended Data Fig. 3a). In naturally apomictic (asexually reproducing) *Pennisetum squamulatum*, an apospory-specific locus contains multiple copies of a *BABY BOOM*-like gene that is expressed in egg cells before fertilization and induces parthenogenesis[16,17]. However, it is not known whether the BBM protein from the apomict has evolved novel capability in functional domains and interactions with other factors[16,17], or whether parthenogenesis might simply be a consequence of the expression pattern. To test whether wild-type rice BBM1 could initiate embryo development without fertilization, we ectopically expressed *BBM1* under an *Arabidopsis* egg-cell-specific promoter (*pDD45*)[18] that has previously been shown to confer egg-cell expression in rice[19] (Extended Data Fig. 3b, c). In emasculated flowers,

[1]Department of Plant Biology, University of California, Davis, CA, USA. [2]Innovative Genomics Institute, Berkeley, CA, USA. [3]Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA, USA. [4]Institut Jean-Pierre Bourgin, INRA, AgroParisTech, CNRS, Université Paris-Saclay, Versailles, France. [5]Department of Plant Sciences, University of California, Davis, CA, USA. *e-mail: sundar@ucdavis.edu

© 2019 Springer Nature Limited. All rights reserved.

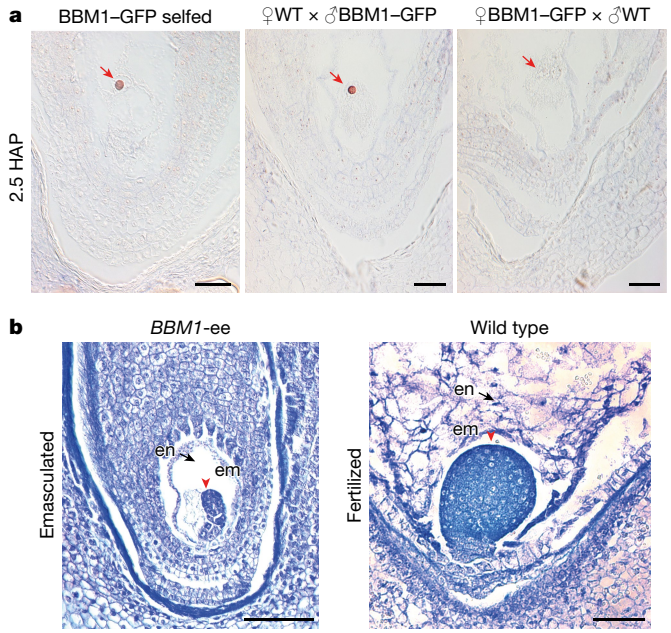


Fig. 1 | Paternal expression of *BBM1* in zygotes. a, Paternal allele-specific expression of *BBM1* in isogenic zygotes at 2.5 HAP. Expression of BBM1 fused to a GFP reporter was detected by antibody staining. GFP expression is observed only when BBM1–GFP is transmitted by the male parent ($n = 20$ for each panel, χ^2 test $P = 0.039$). Left, $n = 11/20$; middle, $n = 9/20$; right, $n = 0/20$. Red arrows point to zygote nuclei. WT, wild type. Scale bars, 25 μm. **b**, Development of parthenogenetic embryos (red arrowhead) by egg-cell-specific expression of *BBM1* in carpels of an emasculated *BBM1*-ee plant at nine days after emasculation ($n = 12/98$). In the absence of fertilization, endosperm development is not observed (black arrow). In fertilized control wild-type (4 days after pollination (DAP)) carpels, the development of both embryo (em; red arrowhead) and endosperm (en; black arrow) is observed ($n = 30$). Scale bars, 100 μm.

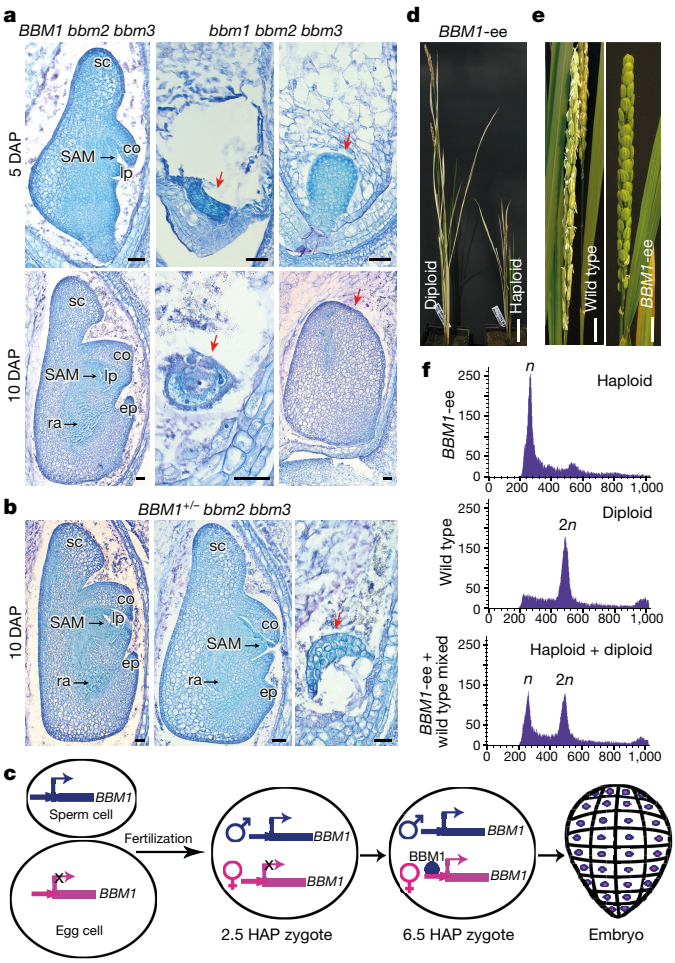


Fig. 2 | Phenotypes of *bbm1 bbm2 bbm3* mutant embryos and haploid induction. a, Embryos at 5 DAP (top) and 10 DAP (bottom). Embryos develop normally with wild-type *BBM1* ($n = 50$; left) but show an early arrest ($n = 24/82$; middle) or undergo a number of divisions without organ formation ($n = 58/82$; right) in *bbm1 bbm2 bbm3* triple homozygous mutant embryos. **b**, 10 DAP embryos that are heterozygous for *BBM1* but homozygous mutants for *bbm2* and *bbm3*. They show normal development ($n = 38/53$, left), are delayed ($n = 8/53$; middle), or show early arrest ($n = 4/53$; right). Scale bars, 100 μm. co, coleoptile; ep, epiblast; lp, leaf primordia; ra, radicle; SAM, shoot apical meristem; sc, scutellum. **c**, Schematic model of *BBM1* function in rice embryogenesis. **d**–**f**, Characterization of *BBM1*-ee induced haploids. **d**, Difference in height between parthenogenetic haploid and sexual diploid siblings ($n = 555$). Scale bar, 5 cm. **e**, A *BBM1*-ee parthenogenetic haploid panicle showing no anthesis (right) compared to an anthesis stage control wild-type panicle (left) ($n = 113$). **f**, Flow-cytometric DNA histograms for ploidy determination. Parthenogenetic haploid showing a 1*n* peak ($n = 19$, top), wild-type diploid with a 2*n* peak (middle) and a mixed sample of *BBM1*-ee and wild type showing 1*n* and 2*n* peaks (bottom).

we observed embryonic structures without endosperm development (Fig. 1b) in around 12% ($n = 98$) of ovules of *pDD45::BBM1* transformants (hereafter referred to as *BBM1*-ee, to denote *BBM1*-egg-cell expressed); these structures were absent in wild-type ovules ($n = 109$). Thus, the expression of a single wild-type transcription factor, BBM1, can overcome the requirement of fertilization for embryo initiation by an egg cell. The observation that a wild-type gene from a sexually reproducing plant is sufficient to induce parthenogenesis when misexpressed suggests that asexual reproduction could potentially evolve from the altered expression of existing genes within the sexual pathway.

Loss-of-function mutants of *BBM*-like genes in *Arabidopsis* and related plants have no embryonic phenotypes; consequently, their functions in early embryogenesis are as yet undefined[2]. Of the multiple *BBM*-like genes in rice, at least three—*BBM1*, *BBM2* and *BBM3* (Os11g19060, Os02g40070 and Os01g67410, respectively)—are consistently expressed in early zygotes (Extended Data Table 1a). We used the CRISPR–Cas9 system to generate *bbm1 bbm3* and *bbm2 bbm3* double mutants (Extended Data Fig. 4a, b), both of which were fully fertile. Crossing the double mutants and selfing (Extended Data Fig. 4c; see Methods) yielded no *bbm1 bbm2 bbm3* triple homozygous plants ($n = 52$). However, *BBM1/bbm1 bbm2/bbm2 bbm3/bbm3* plants were recovered and selfed (Extended Data Fig. 4d). Analysis of the progeny showed that approximately 36% failed to germinate (Extended Data Table 1b). Genotyping of the germinated seedlings suggested that the viability of the *bbm1 bbm2 bbm3* triple-mutant seeds was severely affected (2 out of 191 viable compared with the expected 48 out of 191; Extended Data Table 1b). *BBM1/bbm1 bbm2/bbm2 bbm3/bbm3* seedlings were also under-represented, which suggests that the viability of this genotype is also compromised (Extended Data Table 1b). A subset of the non-germinating seeds could be genotyped using their endosperm, and were found to be either homozygous or heterozygous for *bbm1* but not homozygous for *BBM1* (Extended Data Fig. 4e). The two *bbm1 bbm2 bbm3* triple homozygotes showed normal growth with no obvious vegetative or floral defects and produced normal seed sets, indicating that the *BBM1–BBM3* genes are not required for post-embryonic development. However, their progeny seeds failed to germinate (Extended Data Fig. 4f), confirming the requirement of *BBM1–BBM3* genes for seed viability.

To test whether the parent of origin affects seed viability, we performed reciprocal crosses of *BBM1/bbm1 bbm2/bbm2 bbm3/bbm3* to *BBM1/BBM1 bbm2/bbm2 bbm3/bbm3* plants. When the mutant *bbm1* allele was provided by the male parent, approximately 31% of the *bbm1/BBM1* progeny seeds failed to germinate (Extended Data Table 1c), whereas all progeny germinated when the *bbm1* allele was inherited from the female parent (Extended Data Table 1d). Thus,

© 2019 Springer Nature Limited. All rights reserved.

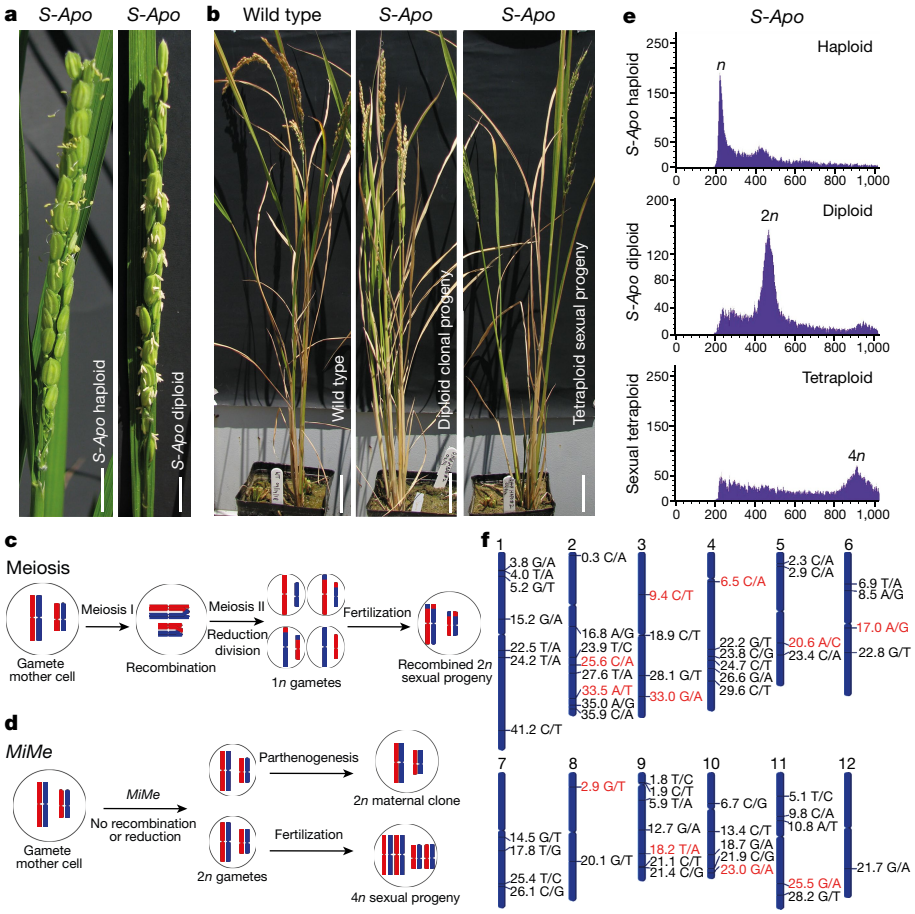


Fig. 3 | Characterization of asexually derived (apomictic) haploids and diploids. a, An *S-Apo* haploid (left; $n = 45$) and *S-Apo* diploid (right; $n = 57$) panicle undergoing anthesis. Scale bars, 1 cm. **b**, Comparison of wild-type (left), *S-Apo* diploid (middle; $n = 57/381$) and sexual tetraploid (right; $n = 324/381$) progeny plants. Scale bars, 5 cm. **c**, **d**, Schematics showing the difference between natural meiosis and *MiMe*. Whereas meiosis and fertilization produce recombined haploid gametes and diploid progeny, *MiMe* leads to the formation of diploid gametes that are clones of the mother plant. Parthenogenesis of a diploid egg cell produces clonal progeny and fertilization of diploid gametes leads to 4*n* sexual progeny. **e**, Flow-cytometric DNA histograms for ploidy determination of *S-Apo* plants. An *S-Apo* haploid (1*n*, top, $n = 30$), an *S-Apo* diploid progeny of a diploid *S-Apo* parent showing a 2*n* peak (middle; $n = 26$) and a sexual tetraploid progeny of a diploid *S-Apo* parent shows a 4*n* peak (bottom; $n = 90$). The *x* axis is the measure of relative fluorescence and the *y* axis shows the number of nuclei. **f**, Chromosomal view showing 57 heterozygous SNPs (position in Mb) identified in the T_0 *S-Apo* mother plant of line 1. The SNPs labelled in red are those additionally confirmed by PCR.

seed viability depends upon a functional *BBM1* allele from the male parent, consistent with male-specific expression of *BBM1* in zygotes. Next we investigated the embryo phenotypes of *bbm2 bbm3* progeny seeds segregating for the *bbm1* mutation. The *bbm1 bbm2 bbm3* embryos were either arrested early or underwent growth by cell division without any corresponding developmental patterning (Fig. 2a). By contrast, embryos that were heterozygous (*BBM1/bbm1 bbm2/bbm2 bbm3/bbm3*) showed a range of phenotypes—from normal to delayed development (Fig. 2b)—as well as the early arrest or unstructured growth phenotypes observed in the triple mutant (Fig. 2b, Extended Data Fig. 4g). This range of phenotypes might occur by partial rescue from late expression of the female *BBM1* allele. Additionally, *BBM4* (Os04g42570)—a fourth *BBM*-like gene that also shows detectable expression in male gametes (Extended Data Table 1a)—might provide sufficient residual function for partial rescue. The recovery of around 0.7% of the *bbm1 bbm2 bbm3* triple homozygous plants is consistent with the hypothesis of residual *BBM* function being provided by *BBM4* (Extended Data Table 1b).

Together, these data suggest that male-genome-derived expression of *BBM1*—acting redundantly with other *BBM* genes—triggers the embryonic program in the fertilized egg cell. Subsequent activation of expression of the female *BBM1* allele by the male BBM1 results in biallelic expression, with both parental alleles eventually contributing to embryo patterning and organ morphogenesis (Fig. 2c). *BBM*-like genes have been shown to promote regeneration from tissue culture, suggesting that they act as pluripotency factors[20]. Our study supports a model in which the requirement of fertilization to initiate embryogenesis in rice arises from the dependency of the zygote on the male gamete for the expression of pluripotency factors after fertilization. This is in contrast to embryogenesis in vertebrate animals, in which pluripotency factors are maternally provided[7]. As demonstrated below, the requirement for fertilization can therefore be bypassed by driving the expression of one such factor from the female gamete.

Haploid plants are efficient tools for the acceleration of plant breeding, because homozygous isogenic lines can be produced in one generation after chromosome doubling[21]. The expression of *BBM1* in the egg cell initiated parthenogenesis in emasculated flowers (Fig. 1b), but the seeds aborted in the absence of endosperm (Extended Data Fig. 3d). Self-pollinated T_1 progeny from *BBM1*-ee transgenic plants were analysed to determine whether endosperm development by fertilization could produce viable seeds containing parthenogenetically derived haploid embryos. We identified haploids by their small size compared with their diploid siblings, as well as by their sterile flowers owing to defective meiosis[22] (Fig. 2d, e, Extended Data Fig. 5a–d). The ploidy of haploid T_1 plants was confirmed by flow cytometry (Fig. 2f). The haploid induction frequency was 5–10% (T_1 plants) and reached around 29% in homozygous T_2 line 8C—this frequency was maintained through multiple generations (Extended Data Table 2a). Thus, misexpression of

© 2019 Springer Nature Limited. All rights reserved.

the wild-type *BBM1* gene in the egg cell is sufficient for the production of haploid plants.

Crop yields can be improved markedly by the use of F_1 hybrid plants that exhibit enhanced vigour ('hybrid vigour'). If meiosis and fertilization are bypassed, hybrids could be propagated through seeds without segregation. Asexual propagation through seeds—known as apomixes—is known to occur naturally in more than 400 species, although not in the major crop plants[23,24]. The development of a method to introduce apomixis into crop plants has been described as 'the holy grail of agriculture'[5] as it can enable fixation of hybrid vigour and stabilization of superior heterozygous genotypes in breeding programs[6,25]. A genetic approach called *MiMe*, which eliminates recombination and substitutes mitosis for meiosis (Fig. 3c, d), has been reported in *Arabidopsis*[3] and rice[4]. In *MiMe*, a triple knockout of the meiotic genes *REC8*, *PAIR1* and *OSD1* produces unrecombined diploid male and female gametes. We tested the possibility that *BBM1*-ee-induced parthenogenesis in rice combined with *MiMe* could result in asexual propagation through seeds (Extended Data Fig. 5f). The three rice *MiMe* genes[4] were subject to genome editing by CRISPR–Cas9 in haploid and diploid plants carrying the *BBM1*-ee transgene (Extended Data Fig. 6a). Unlike *BBM1-ee* haploids, the *MiMe* + *BBM1*-ee haploids were fertile (Extended Data Fig. 6c, d) with normal anther development (Fig. 3a), suggesting that meiosis was successfully replaced by mitosis. Self-pollination of *MiMe* plants invariably results in doubling of the chromosome number[22], so the progeny of haploid *MiMe* plants should be diploid (double haploid). However, we obtained haploid progeny from two *MiMe* + *BBM1*-ee (hereafter denoted *S-Apo*, for *Synthetic-Apomictic*) haploid mother plants at frequencies of 26% and 15%, due to parthenogenesis (Fig. 3e, top, Extended Data Table 2b). These haploid induction frequencies were maintained for the next two generations (Extended Data Table 2b). These results show that haploid *S-Apo* plants can be propagated asexually through seeds. Additionally, the sexual T_1 double-haploid ($2n$) progeny from the haploid *S-Apo* plants yielded both diploid and tetraploid plants in the T_2, T_3 and T_4 generations; the former class is expected from the successful asexual propagation of double haploids (Extended Data Table 2b).

For the clonal propagation of diploid *S-Apo* plants, we obtained two fertile transformants with the requisite six null mutations in three *MiMe* genes (Extended Data Fig. 7a, b). Diploid *MiMe* rice plants have been previously shown—despite reduced seed sets—to produce exclusively tetraploid progeny by sexual reproduction and no diploids[4] (Extended Data Fig. 6c). However, we obtained diploids at frequencies of 11% and 29% (Extended Data Table 2b) from the progeny of two diploid *S-Apo* (that is, *MiMe* + *BBM1*-ee) T_0 transformants (Fig. 3b–e, Extended Data Fig. 6e). The rest of the progeny were tetraploid (Fig. 3e). The progeny of a control *MiMe* diploid plant were all determined to be tetraploid (Extended Data Fig. 6b, c). Because T_1 diploid progeny of T_0 diploid *S-Apo* parents are predicted to arise from the parthenogenesis of unreduced female gametes, they should be clonal with the parent and should not exhibit genetic segregation. The T_1 diploids were propagated, and two more generations (T_2 and T_3) of diploid clones were identified by flow cytometry screening.

To demonstrate clonal propagation, we performed whole-genome sequencing on a diploid T_0 *S-Apo* mother plant (line 1), two diploid T_1 progeny, two T_2 diploid progeny of diploid T_1 plants and a control untransformed wild-type plant. Analysis for sequence variants identified 57 heterozygous SNPs in unique sequences distributed over the genome in the T_0 mother plant (Fig. 3f, Supplementary Table 1) that are non-variant in the wild-type plant (see Methods). These 57 SNPs were determined to be heterozygous in all four T_1 and T_2 diploid progeny sequenced. The probability of any single progeny retaining heterozygosity by random segregation for just a subset of 22 unlinked SNPs on different chromosome arms is $P = 2.4 \times 10^{-7}$. The maintenance of heterozygosity at all 57 loci for two generations confirms that the diploid progeny are clonally generated by asexual reproduction. The T_0 *S-Apo* mother (line 1) is additionally biallelic for mutations in the *PAIR1* and *REC8* genes, as were all T_1, T_2 and two T_3 diploid progeny tested (Extended Data Fig. 7a). For SNP validation, 11 randomly selected SNPs were amplified by PCR followed by Sanger sequencing[26] and found to be conserved in the T_0 mother plant and all the T_1, T_2 and T_3 progeny tested (Extended Data Fig. 8). The second diploid *S-Apo* transformant (line 5) is biallelic for all three *MiMe* genes (Extended Data Fig. 7b) and also heterozygous for one of the 11 SNPs confirmed by PCR for line 1. Five T_1 diploid progeny carried an identical set of alleles to the T_0 mother (Extended Data Fig. 7b). The probability that all five progeny would inherit heterozygosity at these four loci by random segregation is $P = 1.8 \times 10^{-5}$. These findings from an independently generated apomictic parent provide further support for successful clonal propagation.

This study demonstrates that asexual propagation without genetic segregation can be engineered in a sexually reproducing plant, and illustrates the feasibility of clonal propagation of hybrids through seeds in rice. Seed formation in this system still requires fertilization to make endosperm (Extended Data Fig. 5f). This endosperm is expected to be hexaploid owing to fertilization of a tetraploid central cell by a diploid sperm cell, whereas the parthenogenetic embryo is diploid, giving a 3:1 ploidy ratio. This deviation from the normal 3:2 ploidy ratio of endosperm to embryo does not appear to be consequential for viability or seed size (Extended Data Fig. 6f, g). Additionally, the clonally propagated seeds preserve the 2:1 maternal-to-paternal genome ratio in endosperm that is required for seed viability[27,28]. To engineer a completely asexual system involving autonomous endosperm formation may not be straightforward in a sexually reproducing crop, and nor is it essential, as many natural apomicts also form seeds with fertilized endosperm[23]. The efficiency of clonal propagation in our system is in part limited by the frequency of parthenogenesis, which could potentially be improved in the future, for example with different promoters. An important factor to consider for future rice-breeding strategies is that genome-wide heterozygosity may be less critical for yield than the incorporation of specific alleles that exhibit full or partial dominance[29,30]. Nevertheless, hybrids can provide a rapid route to higher yields from favourable gene combinations, and have been extensively exploited in maize. Because homologous *BBM*-like and *MiMe* genes are found in other cereal crops, including maize[2,20], the methods described here for asexual propagation through synthetic apomixis should be generally extendible to most cereal crops.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0785-8.

Received: 8 June 2018; Accepted: 29 October 2018;
Published online 12 December 2018.

1. Palovaara, J., de Zeeuw, T. & Weijers, D. Tissue and organ initiation in the plant embryo: a first time for everything. *Annu. Rev. Cell Dev. Biol.* **32**, 47–75 (2016).
2. Horstman, A., Willemsen, V., Boutilier, K. & Heidstra, R. AINTEGUMENTA-LIKE proteins: hubs in a plethora of networks. *Trends Plant Sci.* **19**, 146–157 (2014).
3. d'Erfurth, I. et al. Turning meiosis into mitosis. *PLoS Biol.* **7**, e1000124 (2009).
4. Mieulet, D. et al. Turning rice meiosis into mitosis. *Cell Res.* **26**, 1242–1254 (2016).
5. Sailer, C., Schmid, B. & Grossniklaus, U. Apomixis allows the transgenerational fixation of phenotypes in hybrid plants. *Curr. Biol.* **26**, 331–337 (2016).
6. Vielle Calzada, J.-P., Crane, C. F. & Stelly, D. M. Apomixis—the asexual revolution. *Science* **274**, 1322–1323 (1996).
7. Lee, M. T., Bonneau, A. R. & Giraldez, A. J. Zygotic genome activation during the maternal-to-zygotic transition. *Annu. Rev. Cell Dev. Biol.* **30**, 581–613 (2014).
8. Nodine, M. D. & Bartel, D. P. Maternal and paternal genomes contribute equally to the transcriptome of early plant embryos. *Nature* **482**, 94–97 (2012).
9. Autran, D. et al. Maternal epigenetic pathways control parental contributions to *Arabidopsis* early embryogenesis. *Cell* **145**, 707–719 (2011).
10. Del Toro-De León, G., García-Aguilar, M. & Gillmor, C. S. Non-equivalent contributions of maternal and paternal genomes to early plant embryogenesis. *Nature* **514**, 624–627 (2014).
11. Anderson, S. N. et al. The zygotic transition is initiated in unicellular plant zygotes with asymmetric activation of parental genomes. *Dev. Cell* **43**, 349–358. e4, (2017).
12. Kim, S., Soltis, P. S., Wall, K. & Soltis, D. E. Phylogeny and domain evolution in the *APETALA2*-like gene family. *Mol. Biol. Evol.* **23**, 107–120 (2006).

© 2019 Springer Nature Limited. All rights reserved.

13. Boutilier, K. et al. Ectopic expression of BABY BOOM triggers a conversion from vegetative to embryonic growth. *Plant Cell* **14**, 1737–1749 (2002).
14. Passarinho, P. et al. BABY BOOM target genes provide diverse entry points into cell proliferation and cell growth pathways. *Plant Mol. Biol*. **68**, 225–237 (2008).
15. Anderson, S. N. et al. Transcriptomes of isolated *Oryza sativa* gametes characterized by deep sequencing: evidence for distinct sex-dependent chromatin and epigenetic states before fertilization. *Plant J*. **76**, 729–741 (2013).
16. Conner, J. A., Mookkan, M., Huo, H., Chae, K. & Ozias-Akins, P. A parthenogenesis gene of apomict origin elicits embryo formation from unfertilized eggs in a sexual plant. *Proc. Natl Acad. Sci. USA* **112**, 11205–11210 (2015).
17. Conner, J. A., Podio, M. & Ozias-Akins, P. Haploid embryo production in rice and maize induced by *PsASGR-BBML* transgenes. *Plant Reprod*. **30**, 41–52 (2017).
18. Steffen, J. G., Kang, I. H., Macfarlane, J. & Drews, G. N. Identification of genes expressed in the *Arabidopsis* female gametophyte. *Plant J*. **51**, 281–292 (2007).
19. Ohnishi, Y., Hoshino, R. & Okamoto, T. Dynamics of male and female chromatin during karyogamy in rice zygotes. *Plant Physiol*. **165**, 1533–1543 (2014).
20. Lowe, K. et al. Morphogenic regulators *Baby boom* and *Wuschel* improve monocot transformation. *Plant Cell* **28**, 1998–2015 (2016).
21. Murovec, J. & Bohanec, B. in *Plant Breeding* (ed. Abdurakhmonov, I. Y.) Ch. 5 (IntechOpen, London, 2012).
22. Cifuentes, M., Rivard, M., Pereira, L., Chelysheva, L. & Mercier, R. Haploid meiosis in *Arabidopsis*: double-strand breaks are formed and repaired but without synapsis and crossovers. *PLoS ONE* **8**, e72431 (2013).
23. Hand, M. L. & Koltunow, A. M. The genetic control of apomixis: asexual seed formation. *Genetics* **197**, 441–450 (2014).
24. Ozias-Akins, P. & van Dijk, P. J. Mendelian genetics of apomixis in plants. *Annu. Rev. Genet*. **41**, 509–537 (2007).
25. Khush, G. S. *Apomixis: exploiting hybrid vigor in rice*. (International Rice Research Institute, 1994).
26. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl Acad. Sci. USA* **74**, 5463–5467 (1977).
27. Lafon-Placette, C. & Köhler, C. Endosperm-based postzygotic hybridization barriers: developmental mechanisms and evolutionary drivers. *Mol. Ecol*. **25**, 2620–2629 (2016).
28. Sekine, D. et al. Dissection of two major components of the post-zygotic hybridization barrier in rice endosperm. *Plant J*. **76**, 792–799 (2013).
29. Hua, J. et al. Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc. Natl Acad. Sci. USA* **100**, 2574–2579 (2003).
30. Huang, X. et al. Genomic architecture of heterosis for yield traits in rice. *Nature* **537**, 629–633 (2016).

Acknowledgements We thank U. Vijayraghavan for providing pUN and pUGN vectors; S. Kappu, Z. Liechty and C. Santos-Medellín for advice and help with flow cytometry and sequence analysis; B. Van Bockern for rice transformations; and B. Nguyen and A. Yalda for technical assistance, including genotyping and transplantation. This research was supported by research grants from the National Science Foundation (NSF) (IOS-1547760) and the Innovative Genomics Institute to V.S., NSF grant IOS-1810468 to B.Y., National Institutes of Health grant 1S10OD010786-01to the University of California-Davis Genome Center, and by the United States Department of Agriculture Agricultural Experiment Station (project number CA-D-XXX-6973-H). R.M. acknowledges support from the LabEx Saclay Plant Sciences-SPS (ANR-10-LABX-0040-SPS) to the Institut Jean-Pierre Bourgin.

Reviewer information *Nature* thanks T. Dresselhaus and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions V.S. and I.K. designed the study. I.K. performed experiments and analysed data. D.S. performed analysis of the genome sequences. B.Y. provided pENTR-sgRNA and pUbi-Cas9 vectors for genome editing. V.S. and I. K. wrote the manuscript with input from R.M.

Competing interests The University of California-Davis has filed a patent application on haploid production (PCT/US2017/063249) and a provisional patent application on synthetic apomixis (US62/678,169) arising from this work. INRA has filed a patent application on the use of the MiMe system (EP2208790). The authors declare no other competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0785-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0785-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to V.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Plant materials and growth conditions. Rice cultivar Kitaake (*O. sativa* L. subsp. *japonica*) was used for transformations for raising transgenic lines and as a wild-type control. Wild-type, mutant and transgenic seeds were germinated on half-strength Murashige and Skoog's (MS) medium[31] containing 1% sucrose and 0.3% phytagel in a growth chamber for 12 days, under a 16 h light:8 h dark cycle at 28 °C and 80% relative humidity. Seedlings were then transferred to a greenhouse and grown under natural light conditions in Davis, California.

Chemical treatments. Two-week-old wild-type and *BBM1-GR* seedlings were treated with 0.1% ethanol as mock, 10 μM DEX (Sigma-Aldrich), or 10 μM CYC (Sigma-Aldrich) alone or in combination with 10 μM DEX in liquid half-strength MS[31] salts. Seedlings that were of a similar size and had the same number of leaves were selected for the treatments. Individual biological replicates were constructed using similar leaf samples collected from four different plants, collected for RNA isolation after 24 h. CYC treatments were started 30 min before the DEX treatment in the samples that were treated with both reagents.

Plasmid constructs. Full-length coding sequence (CDS) of *BBM1* was amplified from cDNAs made from rice calli using two sets of primers (KitB1F1 5′-CGGATCCATGGCCTCCATCACC-3′, KitB1R1 5′-CCTTCGACCCCA TCCCAT-3′ and KitB1F2 5′-GGATGGGATGGGGTCGAAG-3′, KitB1R2 3′-GGTACCAGACTGAGAACAGAGGC-3′). The two fragments were fused together by an overlap PCR. The overexpression construct (*BBM1*-ox) was created by cloning *BBM1* coding sequence in pUN vector[32] (Extended Data Fig. 1a). To create the BBM1–GR plasmid (Extended Data Fig. 1e), *BBM1* coding sequence without the stop codon was cloned in pUGN vector[32] for translational fusion with rat glucocorticoid receptor[33]. The whole *BBM1* locus, approximately 3kb upstream sequences and the transcribed region until the stop codon were PCR-amplified in two fragments from genomic DNA using two primer pairs: pB1F1 5′-CTCGAGGTCAACACCAACGCCATC-3′, pB1R1 5′- GAAGTCCTCCAGCTTCGGCGC-3′ and pB1F2 5′-TTGATTGTGTTGATG TGCAGAGTGGGG-3′, pB1R2 5′-CTCGAGCGGTGTCGGCAAAACC-3′. The two fragments were joined at a unique restriction enzyme site, NotI, present downstream of the start codon in the sequence. The whole locus was moved to a pCAMBIA1300 vector already containing *Arabidopsis* histone H2B, eGFP and nopaline synthase gene terminator (Extended Data Fig. 2b). The construct for egg-cell-specific expression of BBM1 was made by cloning *BBM1* downstream to *Arabidopsis* DD45 promoter[18] and upstream of the nopaline synthase terminator (Extended Data Fig. 3b) in pCAMBIA1300.

For genome editing of *BBM1*, *BBM2* and *BBM3* genes, single-guide RNA (sgRNA) sequences 5′-GGAGGACTTCCTCGGCATGC-3′, 5′-GTATGCAATATACTCCTGCC -3′ and 5′-GACGGCGGGAGCTGATCCTG -3′, respectively, were designed by using the web tool https://www.genome.arizona.edu/crispr/ as described[34]. The sgRNAs were cloned in pENTR-sgRNA entry vector. The binary vectors for plant transformations (pCRISPR *BBM1* + *BBM3*, pCRISPR *BBM2* + *BBM3* and pCRISPR *BBM1* + *BBM2* + *BBM3*) were constructed by Gateway LR clonase (Life Technologies) recombination with pUbi–Cas9 destination vector as described[35]. Three candidate genes (*OSD1*, Os02g37850; *PAIR1*, Os03g01590 and *REC8*, Os05g50410) for creating *MiMe* mutations in rice were selected as previously described[4] and sgRNAs sequences 5′-GCGCTCGCCGACCCCTCGGG-3′, 5′-GGTGAG GAGGTTGTCGTCGA-3′ and 5′-GTGTGGCGATCGTGTACGAG-3′, respectively, for CRISPR–Cas9-based knockout were designed as described[34]. Vector pCAMBIA2300 *MiMe* CRISPR–Cas9 (Extended Data Fig. 6a) for plant transformations was constructed as described[35], except the resistance marker in the destination vector pUbi–Cas9 was changed to kanamycin (*Neomycin Phosphotransferase II*). pCAMBIA2300 *MiMe* CRISPR–Cas9 was transformed in embryogenic calli derived from *pDD45::BBM1#8c* haploid inducer lines (Extended Data Fig. 3b). Rice transformations were carried out as previously described[36] at the University of California-Davis plant transformation facility. T_0 plants were grown in a greenhouse and screened for *MiMe* mutations. T_1 plants obtained from seeds were subjected to ploidy determination and genotyping for *MiMe* mutations.

Generating *bbm1 bbm2 bbm3* mutants. Rice embryogenic calli were transformed with pCRISPR *BBM1* + *BBM3*, or pCRISPR *BBM2* + *BBM3*. The transformants that carried the *bbm1 bbm3* and *bbm2 bbm3* double mutations generated by genome editing (Extended Data Fig. 4a, b) did not show any phenotypic abnormalities and were fertile. The two double mutants were crossed and selfed; however, no *bbm1 bbm2 bbm3* triple-homozygous plants were recovered in the F_2 generation (Extended Data Fig. 4c). However, plants heterozygous for *BBM1 (bbm1/BBM1)* but homozygous mutant for both *bbm2* and *bbm3* could be recovered, and their progeny were analysed in detail (Extended Data Fig. 4d).

Genotyping. Genotyping of *BBM1*, *BBM2* and *BBM3* mutants was carried out by PCR-amplifying DNA at the mutation site with primers *BBM1* SeqF 5′-TTGATTGTGTTGATGTGC-3′ *BBM1* SeqR 5′-GAGAGACGACCTACTTG GTGAC-3′; *BBM2* SeqF 5′-TAGCTAGCTTGTTAATAGATCATAG-3′, *BBM2* SeqR 5′-TCATATCTCAGTGTGATAGTCTG-3′; and *BBM3* SeqF 5′-ATGCTGCTGCTCCGAGAAG-3′, *BBM3* SeqR 5′-GCTTAGTGCTCCAAACCTCTC-3′. Sanger sequencing[26] of the three PCR amplicons of 464 bp, 262 bp and 547 bp, respectively, for the three genes was carried out at the University of California-Davis DNA-sequencing facility. Because a 1-bp deletion mutation in *BBM1* disrupted an SphI restriction-enzyme site (Extended Data Fig. 4d), all further genotyping of *BBM1* for mutational analysis was performed with restriction digestion of the PCR amplicon with SphI (Extended Data Fig. 4e). For genotyping developing seeds of 5 DAP onwards, endosperm was used for genotyping and embryos were collected for mutant phenotype analysis. DNA fragments at the mutation sites of three *MiMe* genes were PCR-amplified with primers *OSD1* F 5′-TTACTTGGAAGAGGCAGGAGCC -3′, *OSD1* R 5′-ACCTTGACGACTGACGTGATGTC-3′; *PAIR1* F 5′-GTGG TGTGGTGTGTTCAGGAG-3′, *PAIR1* R 5′-TGGAATCCCCAA TCAGTAAGGCAC-3′; and *REC8* F 5′-GCACTAAGGCTCTCCGGAATTCTC-3′, *REC8* R 5′-AATGGATCAAGGAGGAGGCACC-3′. PCR amplicons of 364 bp, 344 bp and 326 bp—for *OSD1*, *PAIR1* and *REC8*, respectively—were subjected to Sanger sequencing[26] for mutation analysis.

Emasculation, crosses and pollinations. Flowers from *BBM1*-ee T_0 transgenic rice lines were emasculated around the anthesis stage, bagged and allowed to grow for another nine days after emasculation. Carpels were collected and fixed for analysis in formaldehyde (10%)–acetic acid (5%)–ethanol (50%). A translational fusion consisting of the *BBM1* genomic locus to GFP (BBM1–GFP; Extended Data Fig. 2b) was introduced into the inbred *japonica* (Kitaake) cultivar by transformation. Plants hemizygous for the BBM1–GFP transgene were then reciprocally crossed to wild-type plants. Flowers from wild-type or BBM1–GFP transgenic plants were hand-pollinated around the anthesis stage and carpels were collected 2.5 and 6.5 HAP.

For phenotypic analysis of mutant embryos, self-pollinated flowers from mutant plants were scored for anthesis, and collected 5 or 10 DAP. For crosses of *bbm1 bbm3* and *bbm2 bbm3* plants, only T_2 progeny plants in which the CRISPR–Cas9 transgene had already segregated out were used as parents. For all crosses of *bbm1 bbm3* with *bbm2 bbm3* plants, and for the reciprocal crosses between *BBM1/bbm1 bbm2/bbm2 bbm3/bbm3* and *BBM1/BBM1 bbm2/bbm2 bbm3/bbm3* plants, panicles used as females were emasculated and bagged with pollen donor panicles. The bags were gently finger-tapped (twice a day) for the next two days. Male panicles were removed, and female panicles were left bagged to make seeds. F_1 seeds were collected four weeks after pollination.

Immunohistochemistry and toluidine blue staining. Owing to the difficulty of imaging GFP fluorescence in early rice zygotes through the carpel tissue, we used antibodies against GFP to detect zygote expression in sectioned rice carpels. Collected carpels were fixed in formaldehyde (10%)–acetic acid (5%)–ethanol (50%). Tissue embedding and sectioning was performed as described previously[37]. Immunohistochemistry was carried out using standard protocols[38], except an antigen-retrieval step was also included. Antigen retrieval was performed by microwaving the slides in 10 mM sodium citrate buffer (pH 6.0) for 10 min. Rabbit anti-GFP antibody ab6556 (Abcam) was used as the primary antibody and goat anti-rabbit alkaline phosphatase conjugate A9919 (Sigma) was used as the secondary antibody. For toluidine blue staining, after rehydration, sections crosslinked to glass slides were stained with 0.01% toluidine blue for 30 s.

Flow cytometry. Nuclei for fluorescence-activated cell sorting (FACS) analysis were isolated by a leaf-chopping method described previously[39]. The isolated nuclei were stained with propidium iodide at 40 μg ml^{-1} in Galbraith's buffer. FACS analysis and DNA-content estimation was carried out using a Becton Dickinson FACScan system using standard protocols[40,41]. DNA histograms were gated out for the initial debris.

Alexander staining of pollen grains. Stamens were collected just before anthesis. Anthers were put on a glass slide in a drop of Alexander's stain containing 40 μl of glacial acetic acid per millilitre of stain[42]. Anthers were covered with a coverslip and slides were heated at 55 °C on a heating block, until the visible staining of pollen was observed.

Library preparation and sequencing. PCR-free DNA libraries were prepared from a wild-type Kitaake control plant, the T_0 *S-Apo* line 1 mother plant, two T_1 and two T_2 progeny clones from *S-Apo* line 1 with 500 ng of input DNA, using NuGEN Celero DNA-Seq kit, following the manufacturer's instructions. Samples were multiplexed and six libraries per lane were run on Illumina HiSeq platforms at the University of California-Davis Genome Center.

Whole-genome DNA sequencing and statistical analysis. Adaptor removal and quality trimming of 150-bp paired-end reads was performed using Trimmomatic

© 2019 Springer Nature Limited. All rights reserved.

0.38[43] resulting in 13–16 gigabases of sequence for each library. The reads were aligned to the *O. sativa* reference genome (Nipponbare, Release 7.0)[44] using *bwa mem*[45]. To discover variants that were heterozygous in the T_0 mother plant (line 1), the variant finder GATK4.0 HaplotypeCaller was used in single-sample mode[46] and selecting only for SNPs. Repeated elements of the genome were masked from analysis using annotated repeats from http://www.phytozome.org (Osativa_323_v7.0.repeatmasked_assembly_v7.0.gff3). Variants were retained for analysis after filtering on the basis of mapping quality (MQ = 60), QualByDepth (QD >2), StrandOddsRatio (SOR <1.8), unfiltered read depth ($10 \leq DP \leq 40$) and fraction of the alternate allele ($0.4 \leq DP \leq 0.6$), with the expectation that a truly heterozygous locus should show roughly equal numbers of read counts for each allele. To increase certainty that the set of loci included only true heterozygous SNPs, loci which were called heterozygous in the wild-type sample were also discarded. This strategy guards against instances in which incorrect read-mapping over multi-copy regions lead to spurious designation of loci as heterozygous, even though it is likely that we also discarded true heterozygous loci in the process. A final list of 60 high-quality heterozygous SNPs at 57 loci were analysed for segregation in the four progeny clones (T_1 clone A, T_1 clone B, T_2 clone 7 and T_2 clone 21). All SNPs were called heterozygous by HaplotypeCaller in all the progeny samples (Supplementary Table 1).

For statistical analysis of genetic ratios: Either a chi-square goodness-of-fit test or a two-tailed Fisher's exact test was carried out wherever applicable, and the result specified in the legend of the relevant figure or table.

RT–PCR and RT–qPCR. All the cDNAs were synthesized using the iScript cDNA synthesis kit (BioRad) according to the manufacturer's instructions. RT–PCRs were performed with MyTaq Red Mix (Bioline) and RT–qPCRs with iTaq universal SYBR Green supermix (BioRad) using CFX96 Touch real-time PCR system (BioRad). *UBIQUITIN5* (Os03g13170) was used as the internal control and fold changes in the relative abundance of transcripts were calculated as described previously[47]. For RT–qPCR, amplifications for each gene were performed in two biological replicates, and each biological replicate was repeated in three technical replicates for each sample. For *BBM1*, *BBM1* RT F 5′-TACTACCTTTCCGAGGGTTCG-3′ was used in combination with B1RNAi R 5′-GATATC CCAGACTGAGAACAGAGGC -3′ to detect endogenous transcript and with GR RT R 5′-TCTTGTGAGACTCCTGCAGTG-3′ to detect *BBM1-GR* transgenic transcript in RT–qPCR experiments. *BBM1*intronF 5′-GTGGCAGGAAACAAGGATCTG-3′ with B1RNAi R which spanned an intron was used in RT–PCR experiments. For other genes tested in this study, the following primer combinations were used: *LEC1A* F 5′-GACAGGTGATCGAGCTCGTC-3′, *LEC1A* R 5′-CTCTTTCGATGAAACGGTGGC-3′; *LEC1B* F 5′-ACAGC AGCAGAATGGCGATC-3′, *LEC1B* R 5′-CTCATCGATCACTACCTGAACG-3′; *GE* F 5′-CAGGAGCACAAGGCGAAGCG-3′, *GE* R 5′-CTTCGCCTGGATCT CCGGGTG-3′; *OSH1* F 5′-GAGATTGATGCACATGGTGTG-3′, *OSH1* R 5′-CGAGGGGTAAGGCCATTTGTA-3′; and *UBIQUITIN5* F 5′-ACCACTTCGA CCGCCACT-3′, *UBIQUITIN5* R 5′-ACGCCTAAGCCTGCTGGTT-3′.

SNP analysis. Detection of SNPs in *BBM1* transcripts from hybrid zygotes was performed by PCR of 2.5 HAP zygote cDNAs from reciprocally crossed rice *japonica* cultivar Kitaake and *indica* cultivar IR50, as described previously[11]. Primers B1RNAi F 5′-CCTCGAGCAACTATGGTTCGCAGC-3′ and B1RNAi R, which amplified a gene-specific fragment of about 600 bp of *BBM1*, contains 5 SNPs between Kitaake and IR50 (Extended Data Fig. 2a). The PCR amplicons were Sanger-sequenced[26] and chromatograms were analysed for SNPs. For detection of heterozygous SNPs present in the *S-Apo* mother plants and their progeny, 50 ng of input DNA was used for each PCR reaction. Sanger-sequenced[26] PCR chromatograms were analysed for the presence of SNPs. The primers for 11 SNPs analysed are: 1 Chr2 F 5′-TGGGTGCCA CGTTATCTAGG-3′, 1 Chr2 R 5′-GGATTTGGCTACCCTCAAGCT-3′; 2 Chr2 F 5′-GAATGGGCAACTAACAACCGTG-3′, 2 Chr2 R 5′-ACCGTG GAAAGGAACAGCTG-3′; 1 Chr3 F 5′-TGCTGAAGGTGACGTTGATCTG-3′, 1 Chr3 R 5′-CGACGCCAACGAGAAGGA-3′; 2 Chr3 F 5′-GCTCCAGTGCTA GAGAGACATC-3′, 2 Chr3 R 5′-AGCCACCCAGTAACCGTTG-3′; Chr4 F 5′-GATTGGCAAACCAGCTACTGC-3′, Chr4 R 5′-CTGATGGCAAG CTGTTGGC-3′; Chr5 F 5′-ATGATCTGCTGCTTGTTTCAATGC-3′, Chr5 R 5′-TATCCTTCAAGCACCACTGCC-3′; Chr6 F 5′-ACTAATGGGACCACT TGACAGC-3′, Chr6 R 5′-TCAGCCTGAGATGGCTTGG-3′; Chr8 F 5′-CAGACTGTGGGACGCTACATG-3′, Chr8 R 5′-AGAAGATCT GGGCAGCAGTC-3′; Chr9 F 5′-GCTGCACCTGTTAGCTATGTGA-3′, Chr9 R 5′-AGCATCCCAAAAGCACACATG-3′; Chr10 F 5′-TCAGCAGCCTAAGGTT GAAGG-3′, Chr10 R 5′-CTGCTGCTGCTTCATGATCAC-3′; and Chr11 F 5′-GCAGGAACTATTGCCTCTCATGA-3′, Chr11 R 5′-TCAGTCTCATAGCGCA CCAC-3′.

Code availability. Codes for the different analyses are available for non-commercial use from the corresponding author upon request.

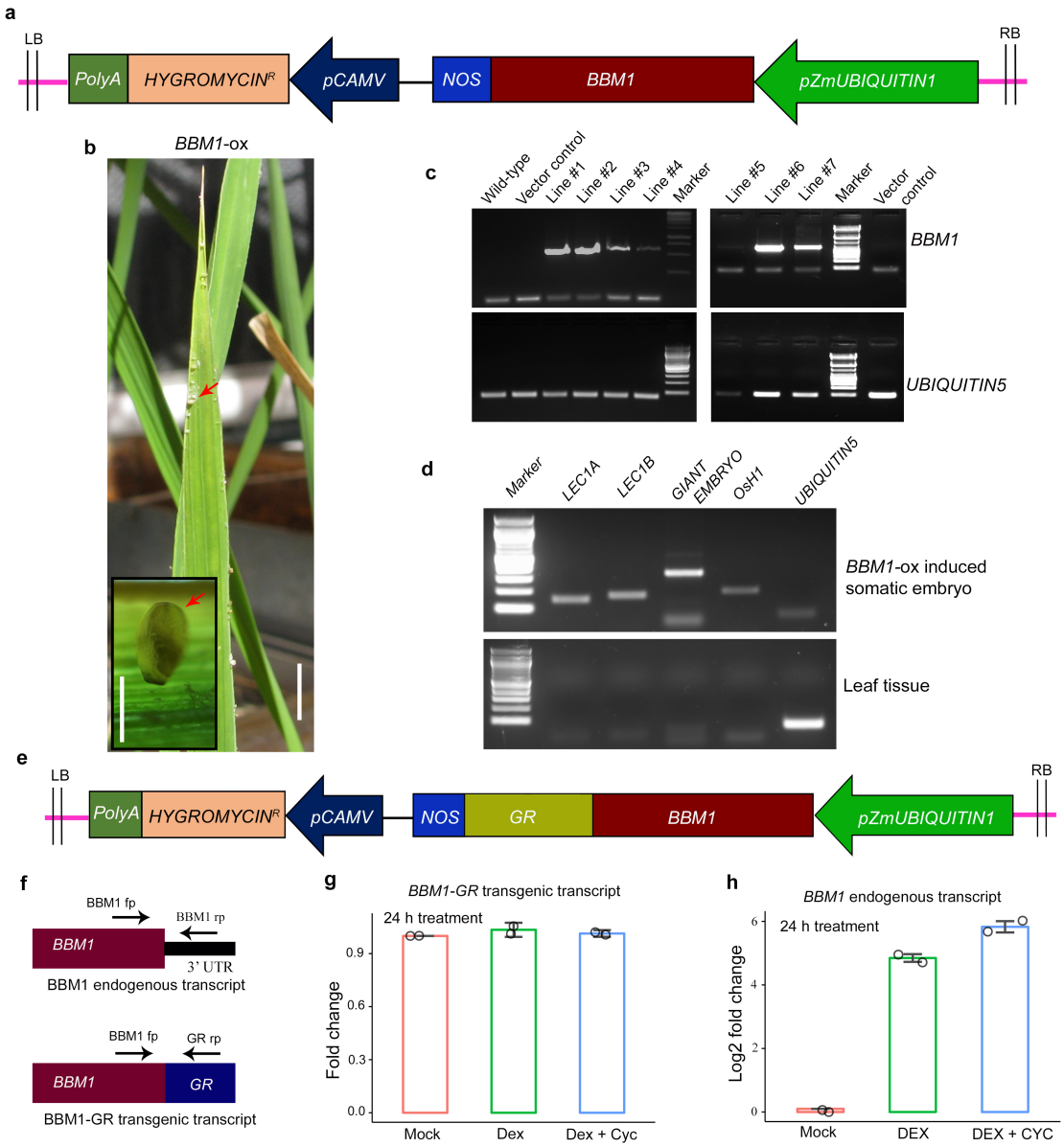
Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Whole-genome DNA sequencing data for *S-Apo* line 1 mother plant, the four progeny clones from two generations, and the Kitaake wild-type control are available from National Center for Biotechnology Information (NCBI) BioProject number PRJNA496208. RNA sequencing data from previously published datasets[11,15] are available from the NCBI Short Read Archive as Project SRP119200 and from the NCBI Gene Expression Omnibus under accession number GSE50777.

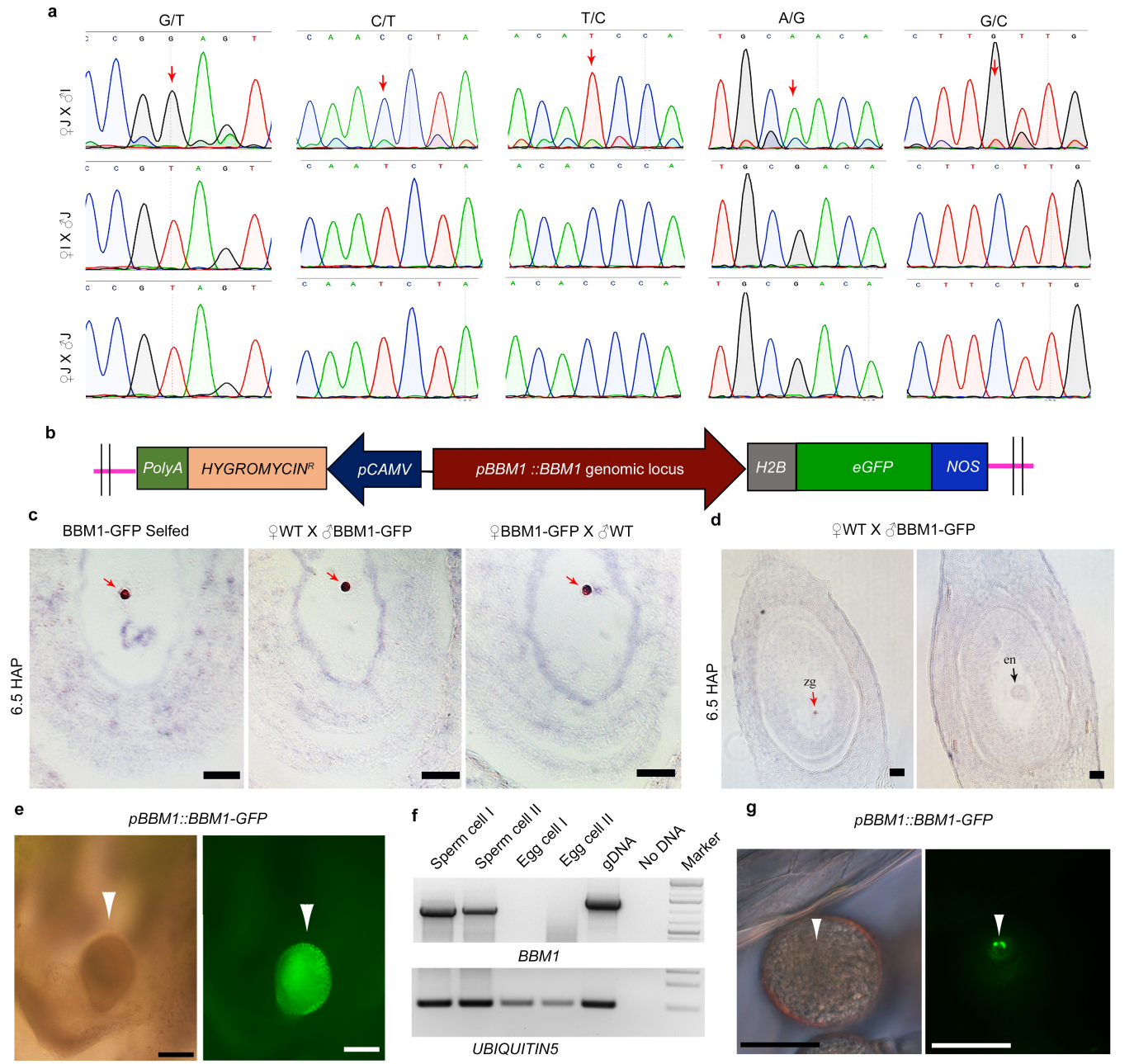
31. Murashige, T. & Skoog, F. A revised medium for rapid growth and bio assays with tobacco tissue cultures. *Physiol. Plant.* **15**, 473–497 (1962).
32. Khanday, I., Yadav, S. R. & Vijayraghavan, U. Rice *LHS1/OsMADS1* controls floret meristem specification by coordinated regulation of transcription factors and hormone signaling pathways. *Plant Physiol.* **161**, 1970–1983 (2013).
33. Aoyama, T. & Chua, N. H. A glucocorticoid-mediated transcriptional induction system in transgenic plants. *Plant J.* **11**, 605–612 (1997).
34. Xie, K., Zhang, J. & Yang, Y. Genome-wide prediction of highly specific guide RNA spacers for CRISPR–Cas9-mediated genome editing in model plants and major crops. *Mol. Plant* **7**, 923–926 (2014).
35. Zhou, H., Liu, B., Weeks, D. P., Spalding, M. H. & Yang, B. Large chromosomal deletions and heritable small genetic changes induced by CRISPR/Cas9 in rice. *Nucleic Acids Res.* **42**, 10903–10914 (2014).
36. Hiei, Y. & Komari, T. Agrobacterium-mediated transformation of rice using immature embryos or calli induced from mature seed. *Nat. Protoc.* **3**, 824–834 (2008).
37. Javelle, M., Marco, C. F. & Timmermans, M. In situ hybridization for the precise localization of transcripts in plants. *J. Vis. Exp.* **57**, e3328 (2011).
38. Sessions, A. Immunohistochemistry on sections of plant tissues using alkaline-phosphatase-coupled secondary antibody. *Cold Spring Harb. Protoc.* https://doi.org/0.1101/pdb.prot4946 (2008).
39. Galbraith, D. W. et al. Rapid flow cytometric analysis of the cell cycle in intact plant tissues. *Science* **220**, 1049–1051 (1983).
40. Doležel, J., Greilhuber, J. & Suda, J. Estimation of nuclear DNA content in plants using flow cytometry. *Nat. Protoc.* **2**, 2233–2244 (2007).
41. Cousin, A., Heel, K., Cowling, W. A. & Nelson, M. N. An efficient high-throughput flow cytometric method for estimating DNA ploidy level in plants. *Cytometry A* **75A**, 1015–1019 (2009).
42. Alexander, M. P. Differential staining of aborted and nonaborted pollen. *Stain Technol.* **44**, 117–122 (1969).
43. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
44. Kawahara, Y. et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (N. Y.)* **6**, 4 (2013).
45. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
46. Van der Auwera, G. A. et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **11**, 11.10.1–11.10.33 (2013).
47. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta CT}$ method. *Methods* **25**, 402–408 (2001).

© 2019 Springer Nature Limited. All rights reserved.



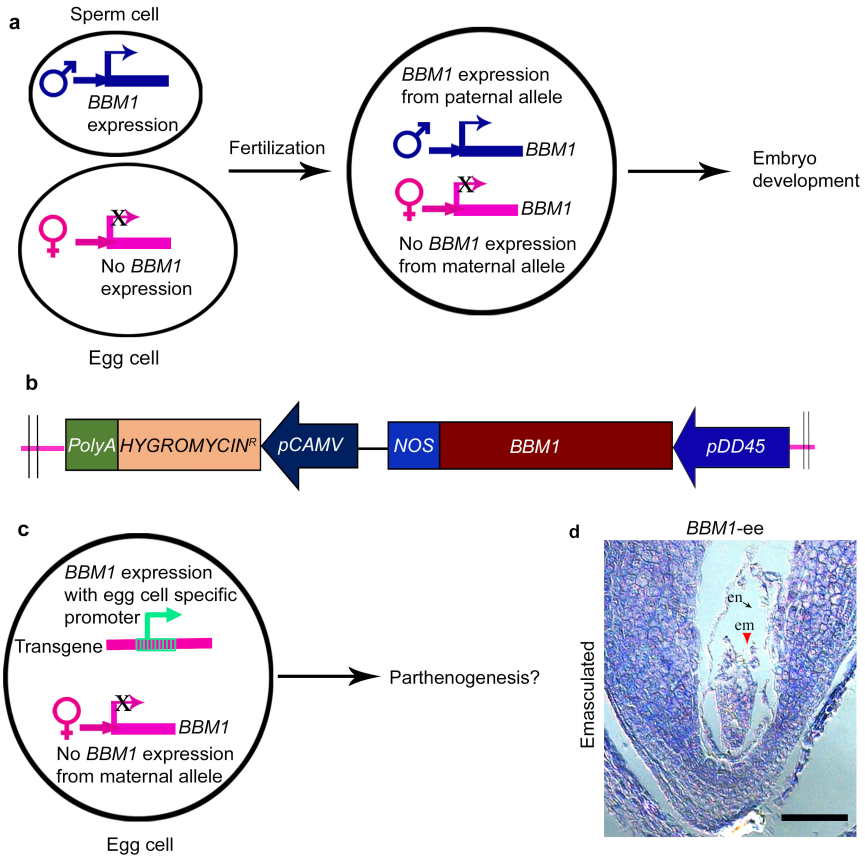
Extended Data Fig. 1 | *BBM1*-induced somatic embryogenesis and auto-activation. a, Schematic of binary construct between T-DNA borders used for ectopic expression (*BBM1*-ox). **b**, Somatic embryo-like structures induced by *BBM1* ectopic expression in rice leaves ($n = 14/20$ transgenic lines). Scale bar, 1 cm. Inset, magnified view of a somatic embryo; scale bar, 0.5 mm. Fourteen of the twenty transgenic plants raised showed the development of such embryo-like structures observed on adult seedlings from the fourth leaf onwards. **c**, Confirmation by RT–PCR of ectopic *BBM1* expression in leaf tissues of transgenic lines. *BBM1* is not expressed in wild-type leaves ($n = 2$ independent replicates). **d**, RT–PCR of embryo marker genes to confirm the embryo identity of somatic embryos induced by *BBM1* overexpression. *OsH1*, *O. sativa HOMEOBOX1*; *LEC1*, *LEAFY COTYLEDON1* ($n = 2$ independent biological replicates). **e**, Schematic of plasmid construct for DEX-inducible BBM1–GR expression system. **f**, Schematic showing primer combinations to distinguish between endogenous *BBM1* and *BBM1-GR* fusion transcripts. **g**, RT–qPCR for fold changes in *BBM1-GR* fusion transcript in samples treated for 24 h with the indicated reagents, showing essentially no differences between treatments. $n = 2$ independent biological replicates (see Methods), data are mean ± s.e.m. and each data point represents the average fold change from three replicates. **h**, Autoactivation of *BBM1* in samples treated with DEX for 24 h, detected by RT–qPCR. $n = 2$ independent biological replicates (see Methods), data are mean ± s.e.m. and each data point represents the average fold change (measured as $\log_2$(change in expression)) from three replicates.

© 2019 Springer Nature Limited. All rights reserved.



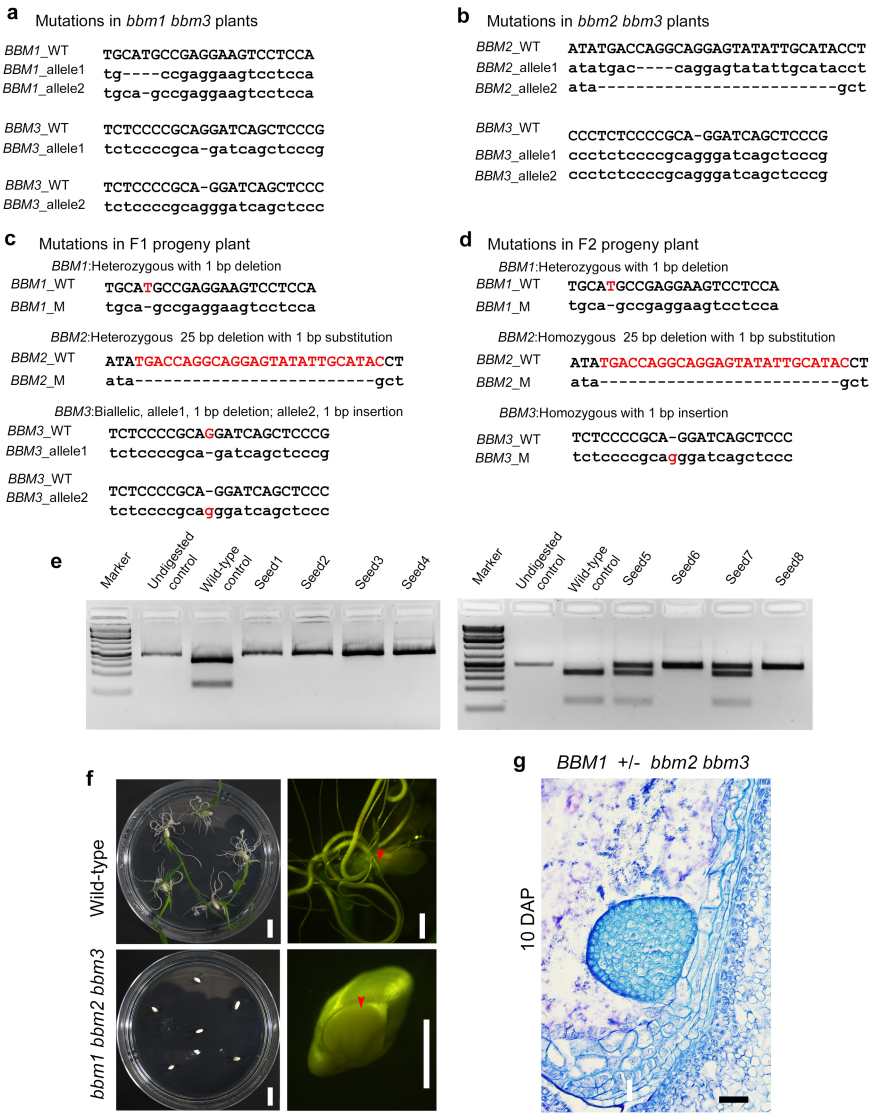
Extended Data Fig. 2 | *BBM1* expression in zygotes and gametes. **a**, Five SNPs sequenced after RT–PCR amplification (red arrows), showing expression only from the male allele in hybrid (J, *japonica*; I, *indica*) 2.5 HAP zygotes ($n = 2$ biological replicates). **b**, Schematic of the *BBM1-GFP* binary construct. **c**, Immunohistochemistry showing expression from both male and female *BBM1* alleles in isogenic 6.5 HAP zygote nuclei ($n = 20$), as compared to male-specific expression at 2.5 HAP (Fig. 1a). Scale bars, 25 μm. **d**, Holistic view of a 6.5 HAP embryo sac showing BBM1–GFP expression in the zygote nucleus (left), while in the same embryo sac expression is not detected in the dividing endosperm (right). zg, zygote. $n = 20$. Scale bar 100 μm. **e**, BBM1–GFP expression in globular-stage rice embryos (white arrowhead, $n = 30$). Differential interference contrast image (left); fluorescence image (right panel). Scale bars, 200 μm. **f**, RT–PCR showing *BBM1* expression in sperm cells; however, the transcript is not detected in egg cells ($n = 2$ independent biological replicates). Primers used for detecting *BBM1* transcript span an intron (see Methods). **g**, BBM1–GFP expression in sperm cells (white arrowhead points to sperm nuclei, $n = 20$). Differential interference contrast image (left) and fluorescent image (right) of a germinating pollen grain showing BBM1–GFP expression in the two sperm cell nuclei.

© 2019 Springer Nature Limited. All rights reserved.



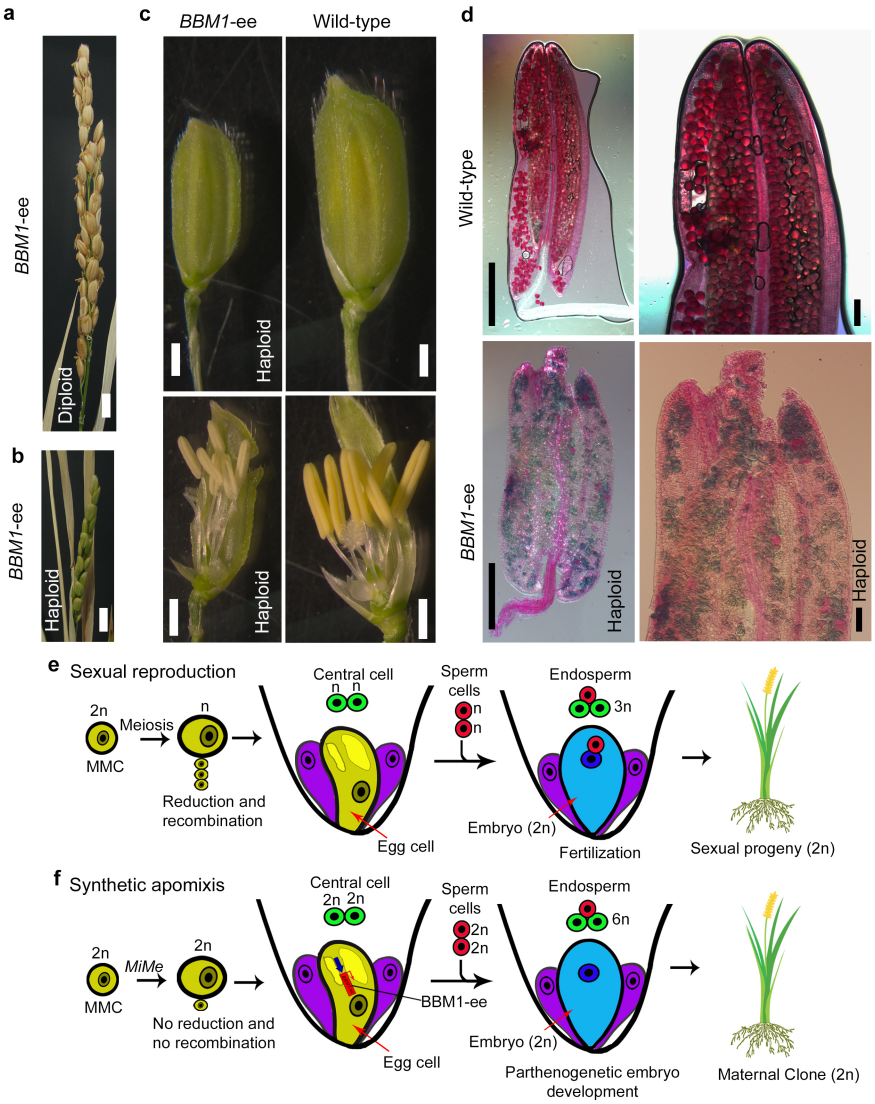
Extended Data Fig. 3 | Parthenogenesis induction by expression of *BBM1* in the egg cell. a, Schematic showing wild-type expression pattern of *BBM1*. **b**, Sketch of T-DNA region of the binary vector used for *BBM1* expression in the egg cell. **c**, Schematic representation of the hypothesis that the expression of *BBM1* in the egg cell can induce parthenogenesis. **d**, A degenerating parthenogenetic embryo (*BBM1*-ee) at 9 days after emasculation (red arrowhead). No endosperm development (black arrow) is observed in emasculated carpels, leading to the abortion of embryos ($n = 12/98$). Scale bar, 100 μm.

© 2019 Springer Nature Limited. All rights reserved.



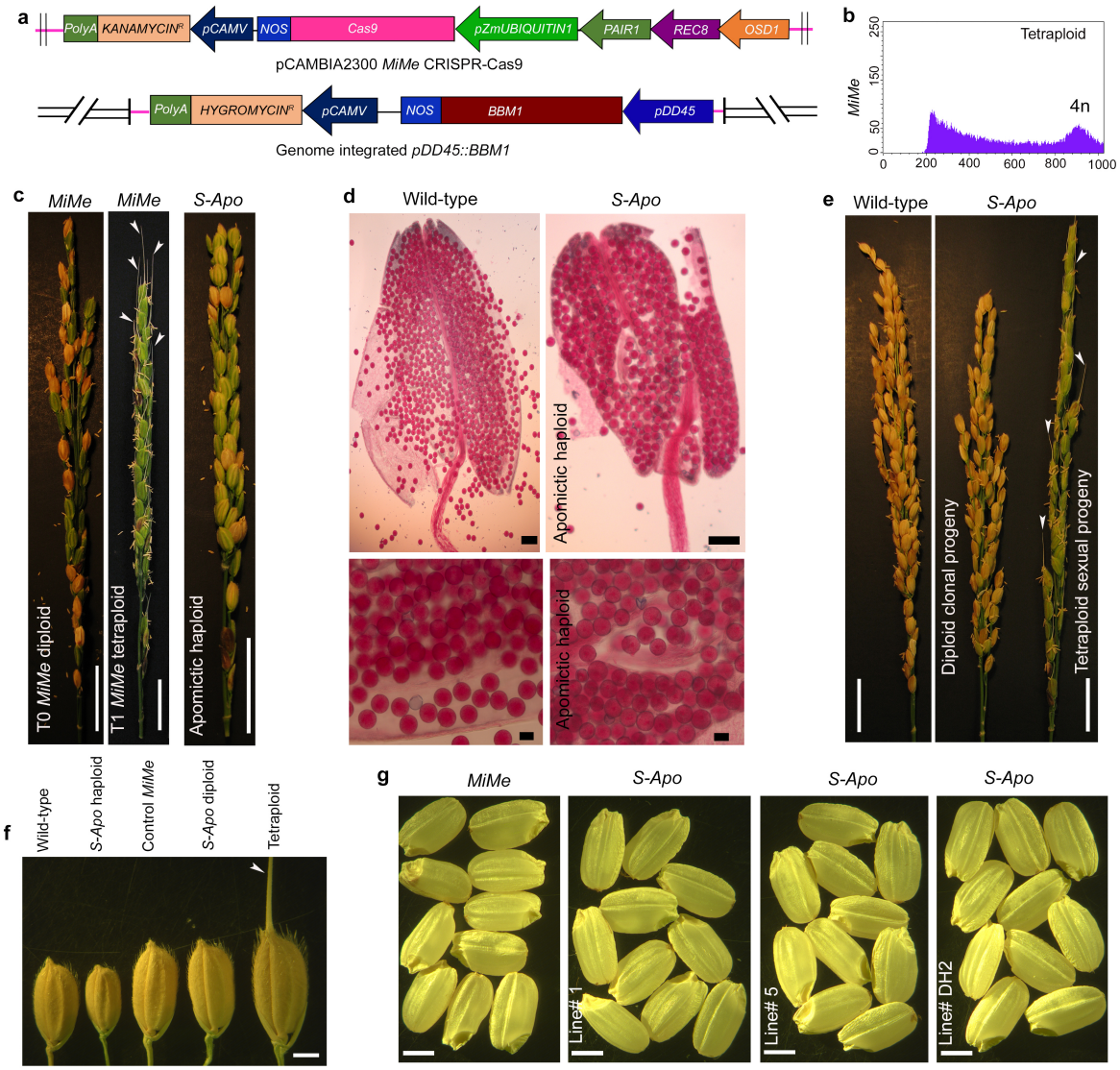
Extended Data Fig. 4 | CRISPR–Cas9 edited mutations in *BBM1*, *BBM2* and *BBM3* in rice. a, DNA sequences of mutations in *bbm1/bbm1 bbm3/bbm3* plants. **b**, DNA sequences of mutations in *bbm2/bbm2 bbm3/bbm3* plants. **a** and **b** were chosen as parents for crosses to generate the *bbm1 bbm2 bbm3* triple homozygous mutants shown in **c** and **d**. **c**, Mutations in the F_1 progeny plant. It is heterozygous for *BBM1* and *BBM2*, and biallelic for *BBM3*. **d**, Mutations in the F_2 progeny plant used for genetic analysis. The plant is heterozygous for *BBM1* with a 1-bp deletion. The *BBM2* locus has a homozygous 25-bp deletion and 1-bp substitution, and the *BBM3* locus is a homozygous mutant with 1-bp insertion. **e**, Genotyping of non-germinating seeds ($n = 8$). The 1-bp deletion mutation in *BBM1* results in disruption of an SphI restriction site. **f**, Seed lethality in *bbm1 bbm2 bbm3* triple homozygous plants. Top, germinating one-week-old wild-type seeds ($n = 30$). Scale bars, 1 cm. A magnified view is shown on the right. Bottom, non-germinating seeds of *bbm1 bbm2 bbm3* triple homozygous plants ($n = 70$). A zoomed-in image of a non-germinating *bbm1 bbm2 bbm3* seed, one week after plating, is shown on the bottom right. No seedling emerged from the embryo site (red arrowhead). **g**, Additional image of a *BBM1/bbm1* heterozygous *bbm2/bbm2 bbm3/bbm3* homozygous 10 DAP embryo ($n = 3/53$) showing no organ formation, similar to triple homozygote phenotype (see Fig. 2a). Scale bar, 100 μm.

© 2019 Springer Nature Limited. All rights reserved.



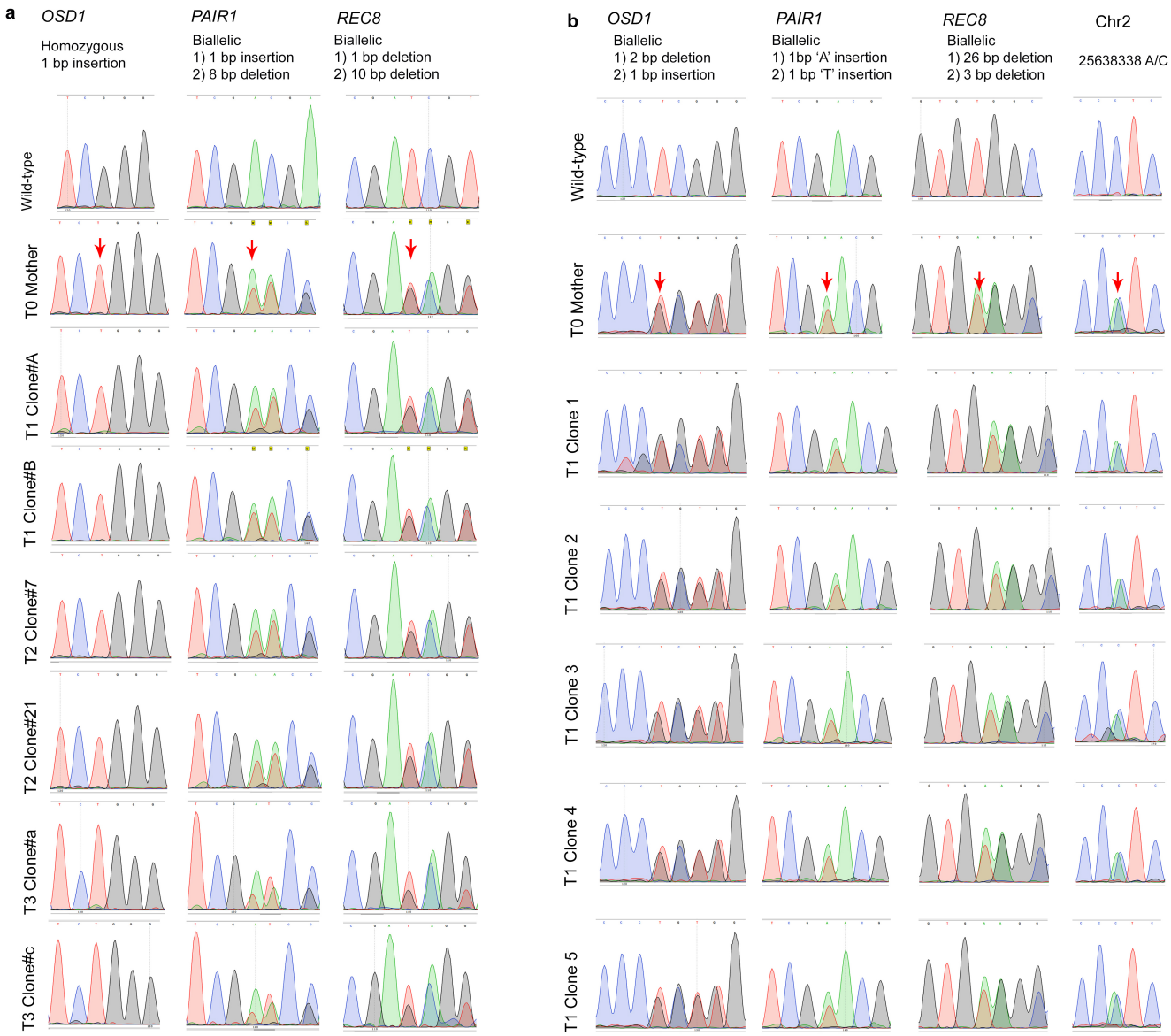
Extended Data Fig. 5 | Haploid induction and synthetic apomixis. Haploids shown are derived from *BBM1*-ee diploids by parthenogenesis. **a**, A control diploid sibling panicle with fertile florets ($n = 442$ plants). Scale bar, 1 cm. **b**, A haploid panicle with infertile florets ($n = 113$ plants). Scale bar, 1 cm. **c**, Differences in floret and floral organ sizes between haploid and control diploid. Left, *BBM1*-ee haploid; right, wild-type control ($n = 20$). Scale bars, 1 mm. **d**, Pollen viability in haploids as assessed by Alexander staining. Top, control wild-type anther with viable pollen ($n = 10$). Bottom, *BBM1*-ee haploid anther with non-viable pollen ($n = 20$). Scale bars, 0.5 mm (left) and 200 μm (right). **e**, **f**, Sexual reproduction compared with asexual reproduction through seed (synthetic apomixis). **e**, Schematic representation of sexual reproduction. Gametes form by meiotic recombination and division; fertilization and gamete fusion give rise to diploid progeny. **f**, Synthetic apomixis. *MiMe* omits meiosis and gives an unrecombined and unreduced ($2n$) egg cell. The $2n$ egg cell is converted parthenogenetically into a clonal embryo by *BBM1*-ee. The endosperm forms in both pathways by fertilization of central cell (homodiploid in wild type, tetraploid in synthetic apomicts) by a sperm cell (haploid in wild type, diploid in synthetic apomicts). The maternal:paternal genome ratio of 2:1 is maintained in the endosperm in both the pathways, ensuring normal seed development.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Asexual propagation through seed in rice. **a**, Top, schematic of the CRISPR–Cas9 plasmid construct used for genome editing of the three *MiMe* rice genes. Bottom, schematic of genome-integrated *pDD45::BBM1* in the *BBM1*-ee plants. **b**, DNA histogram of flow cytometric peak showing 4*n* ploidy in T_1 progeny ($n = 33/33$ tested) of a control T_0 *MiMe* plant. **c**, Left, panicle of a control T_0 diploid *MiMe* plant with fertile seeds. Middle, a tetraploid T_1 *MiMe* panicle, exhibiting complete infertility; that is, no seed filling, and larger flowers (note scale bars), with awns (white arrowhead). Awns are normally suppressed in most *japonica* rice cultivars including Kitaake. All T_1 *MiMe* progeny ($n = 139$) were scored for the phenotype of complete infertility and presence of awns, including 33 plants that were additionally confirmed in **b** by flow cytometry. Right, panicle of an *S-Apo* haploid plant showing fertile seeds ($n = 45$). Scale bars, 2 cm. **d**, Wild-type and *S-Apo* haploid anthers, showing viable pollen ($n = 15$). Scale bars, 0.2 mm (top) and 100 μm (bottom). **e**, Comparison of panicles from wild type (left), with diploid clonal progeny (57/381) and sexual tetraploid progeny ($n = 324/381$) from a diploid *S-Apo* plant (right). The white arrowheads show awns in tetraploid. Scale bars, 2 cm. **f**, Size comparison of progeny seeds from control wild type, a synthetic *S-Apo* haploid, a control *MiMe*, a synthetic *S-Apo* diploid clone, and an infrequent (3%) filled seed produced by the sexual tetraploid progeny of an *S-Apo* diploid ($n = 100$ for each genotype). Scale bar, 2 mm. **g**, Comparison of seed size between control *MiMe*, diploid *S-Apo* line 1, diploid *S-Apo* line 5 and double-haploid *S-Apo* line DH2 ($n = 100$ for each transgenic line). No noticeable variation in seed size is observed. Scale bars, 2 mm.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | *MiMe* mutations and confirmation of clonal progeny from *S-Apo* plants. a, Sequence chromatograms at mutation sites of *MiMe* genes in wild-type, T_0 diploid *S-Apo* mother plant and two diploid progeny from each of T_1, T_2 and T_3 generations of *S-Apo* line 1 ($n = 7$). Red arrows point to mutation sites. *PAIR1* and *REC8* are biallelic whereas *OSD1* is homozygous. **b**, Sequences of the T_0 *S-Apo* mother plant and five T_1 *S-Apo* diploid progeny at *MiMe* mutation sites and one heterozygous SNP in apomixis line 5 ($n = 6$). Red arrows show the mutation sites or SNP. All three *MiMe* mutations—*OSD1*, *PAIR1* and *REC8*—are biallelic. All progeny across different generations in both the *S-Apo* lines have same mutations as the T_0 mother plants, indicating absence of segregation and thus clonal propagation.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Confirmation of SNPs by PCR. Sequence chromatograms of 11 SNPs are shown for wild-type, T_0 diploid *S-Apo* mother plant and two diploid *S-Apo* progeny from each of the T_1, T_2 and T_3 generations for line 1 ($n = 7$). All the 11 SNPs were found to be present in the T_0 mother plant and all the progeny across different generations, confirming that there is no segregation; thus clonal propagation. The red arrows show the location of the SNP. Chr, chromosome; the numbers indicate the position on the respective chromosome.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Functional characterization of *BBM* genes in rice

a

Locus/Isoform	Gene	EC	SpC	Z2.5	Z5	Z9	Z2.5_JxI	Z2.5_IxJ
LOC_Os11g19060.1	*BBM1*	0	2.03	1.45	4.5	5.5	0.51	0.38
LOC_Os02g40070.1	*BBM2*	0	0	0.63	1.75	1.56	0.2	0.28
LOC_Os01g67410.1	*BBM3*	0	0.54	1.01	2.04	0.45	0.15	0.14
LOC_Os04g42570.1	*BBM4*	0	0.2	0	0	0	0.295	0

b

Number of seeds tested	Seeds germinated	Seeds that did not germinate	Percentage non-germinated	Genotypes of germinated seedlings: *BBM1/BBM1: BBM1/bbm1: bbm1/bbm1*
297	191	106	35.6	81:108:2*

c

Female *BBM1/BBM1 bbm2/bbm2 bbm3/bbm3* X Male *bbm1/BBM1 bbm2/bbm2 bbm3/bbm3*					
No. of Seeds	**Seeds germinated**	**Wild-type for *BBM1***	**Heterozygous for *BBM1***	**Seeds did not germinate**	**Non -germinating seeds genotyped**
149	121	59	62	28	23, all heterozygous**

d

Female *bbm1/BBM1 bbm2/bbm2 bbm3/bbm3* X Male *BBM1/BBM1 bbm2/bbm2 bbm3/bbm3*					
No. of Seeds	**Seeds germinated**	**Wild-type for *BBM1***	**Heterozygous for *BBM1***	**Seeds did not germinate**	**Non-germinating seeds genotyped**
67	67	35	32	0	0

a, Expression of four *BBM*-like genes in rice gametes and zygotes from previous studies[11,15] presented as reads per million averaged from three replicates. Z2.5, Z5 and Z9 columns are from isogenic *japonica* zygotes at 2.5, 5 and 9 HAP, respectively. J×I and I×J columns are hybrid zygotes from crosses, the female parent is listed first. EC, egg cell; I, *indica*; J, *japonica*; SpC, sperm cell; Z, zygote. **b**, Summary of seed viability in progeny of *BBM1/ bbm1 bbm2/bbm2 bbm3/bbm3* mutant plants. A loss of viability was observed, as around 36% (106/297) of seeds fail to germinate. Of the germinated seedlings, only 1% (2/191) were triple homozygotes, instead of the expected 25% if there is no effect of genotype on viability. **c**, **d**, Dependence of seed viability on paternal allele transmission of *BBM1*. **c**, When the *bbm1* allele is transmitted by the male parent, around 27% of the genotyped heterozygotes fail to germinate (23/(23 + 62)), despite a functional *BBM1* allele inherited from the female parent. **d**, All seeds germinate when the mutant *bbm1* allele is transmitted by the female parent ($n = 67$).
*The chi-square value for goodness-of-fit between the expected Mendelian 1:2:1 ratio and the observed data is 68.623; the corresponding right-tail P value is 1.714×10^{-15}.
**The two-tailed Fisher's exact test P value is 0.0001, for the genotyped non-germinating seeds to contain all heterozygotes and no wild types.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Haploid induction and clonal propagation in rice

a

Transgenic line#	Generation	Number of plants tested	Number of haploids	% Haploid induction
1	T1	28	2	7.1
3	T1	25	2	8
4	T1	32	3	9.3
5	T1	34	2	5.8
8	T1	57	6	10.5
10	T1	27	2	7.4
11	T1	31	2	6.4
Haploid induction in homozygous T1 progeny line#8c				
8c	T2	185	54	29.2
	T3	40	13	32.5
	T4	33	9	27.2
	T5	18	5	27.7
	T6	21	6	28.5
	T7	24	7	29.1

b

Frequencies of haploid asexual progeny from haploid *S-Apo* plants					
Transgenic line#	**Generation**	**Number of plants tested**	**Number of haploids**	**Number of diploids**	**% Apomixis**
1	T1	19	5	14	26.3
	T2	31	8	23	25.8
2	T1	56	8	47	14.2
	T2	116	19	97	16.3
	T3	34	5	29	14.7
Frequencies of diploid asexual progeny from diploid *S-Apo* plants					
Transgenic line#	**Generation**	**Number of plants tested**	**Number of diploids**	**Number of tetraploids**	**% Apomixis**
1	T1	27	3	24	11.1
	T2	27	4	23	14.8
	T3	13	2	11	15.3
5	T1	41	12	29	29.2
DH#2	T2	121	14	107	11.5
	T3	123	18	105	14.6
	T4	29	4	25	13.7

a, Haploid induction in *BBM1*-ee (*pDD45::BBM1*) transgenic plants. The T_0 primary transformants were hemizygous for the *BBM1*-ee transgene. One diploid T_1 plant 8c from transformant 8 was maintained as a haploid inducer line up to the T_7 generation. **b**, Identification of synthetic haploid and diploid apomictic progeny from *S-Apo* (*MiMe* + *BBM1*-ee) plants of transformant line numbers 1 and 2 (haploids), and line numbers 1 and 5 (diploids). For T_2 and subsequent generations, propagation was performed by selecting from each generation, haploid and diploid progeny respectively. DH#2 refers to a doubled haploid derived from self-pollination of T_1 plants of the haploid apomixis line 2.

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0749-z

Metabolic reprogramming by the *S*-nitroso-CoA reductase system protects against kidney injury

Hua-Lin Zhou[1], Rongli Zhang[1,9], Puneet Anand[1,9], Colin T. Stomberski[1], Zhaoxia Qian[1], Alfred Hausladen[1], Liwen Wang[2], Eugene P. Rhee[3,4], Samir M. Parikh[5,6], S. Ananth Karumanchi[5,6,7] & Jonathan S. Stamler[1,8]*

Endothelial nitric oxide synthase (eNOS) is protective against kidney injury, but the molecular mechanisms of this protection are poorly understood[1,2]. Nitric oxide-based cellular signalling is generally mediated by protein *S*-nitrosylation, the oxidative modification of Cys residues to form *S*-nitrosothiols (SNOs). *S*-nitrosylation regulates proteins in all functional classes, and is controlled by enzymatic machinery that includes *S*-nitrosylases and denitrosylases, which add and remove SNO from proteins, respectively[3,4]. In *Saccharomyces cerevisiae*, the classic metabolic intermediate co-enzyme A (CoA) serves as an endogenous source of SNOs through its conjugation with nitric oxide to form *S*-nitroso-CoA (SNO-CoA), and *S*-nitrosylation of proteins by SNO-CoA is governed by its cognate denitrosylase, SNO-CoA reductase (SCoR)[5]. Mammals possess a functional homologue of yeast SCoR, an aldo-keto reductase family member (AKR1A1)[5] with an unknown physiological role. Here we report that the SNO-CoA–AKR1A1 system is highly expressed in renal proximal tubules, where it transduces the activity of eNOS in reprogramming intermediary metabolism, thereby protecting kidneys against acute kidney injury. Specifically, deletion of *Akr1a1* in mice to reduce SCoR activity increased protein *S*-nitrosylation, protected against acute kidney injury and improved survival, whereas this protection was lost when *Enos* (also known as *Nos3*) was also deleted. Metabolic profiling coupled with unbiased mass spectrometry-based SNO-protein identification revealed that protection by the SNO-CoA–SCoR system is mediated by inhibitory *S*-nitrosylation of pyruvate kinase M2 (PKM2) through a novel locus of regulation, thereby balancing fuel utilization (through glycolysis) with redox protection (through the pentose phosphate shunt). Targeted deletion of PKM2 from mouse proximal tubules recapitulated precisely the protective and mechanistic effects of *S*-nitrosylation in $Akr1a1^{-/-}$ mice, whereas Cys-mutant PKM2, which is refractory to *S*-nitrosylation, negated SNO-CoA bioactivity. Our results identify a physiological function of the SNO-CoA–SCoR system in mammals, describe new regulation of renal metabolism and of PKM2 in differentiated tissues, and offer a novel perspective on kidney injury with therapeutic implications.

SCoR denitrosylases mediate CoA-dependent denitrosylation of proteins (Extended Data Fig. 1a, b), but their role in mammals is unknown. We found that SCoR (also known as AKR1A1, formally an aldoketoreductase of unknown function) is expressed widely, but most abundantly in proximal tubules (Fig. 1a, b). Notably, AKR1A1 constitutes 0.11% of protein in bovine kidney (Extended Data Fig. 1c). eNOS is also expressed highly in proximal tubule epithelial cells, and its expression is induced by acute kidney injury (AKI), whereas neuronal and inducible NO synthase (nNOS and iNOS, respectively) are barely detectable[1,6] (Extended Data Fig. 1d–f). To investigate the physiological role of the SNO-CoA–SCoR system, we created $Akr1a1^{-/-}$ and $Akr1a1^{-/-}Enos^{-/-}$ mice (Extended Data Fig. 1g, h). Activity to metabolize SNO-CoA was markedly reduced in the kidneys of both $Akr1a1^{-/-}$ and $Akr1a1^{-/-}Enos^{-/-}$ mice (Fig. 1c, d).

We subjected wild-type and $Akr1a1^{-/-}$ mice to ischaemia–reperfusion (I/R)-induced AKI. Notably, activity to metabolize SNO-CoA was reduced after AKI in wild-type mice (Extended Data Fig. 2a–c). Serum creatinine and blood urea nitrogen (BUN), indicators of kidney dysfunction, were significantly lower in $Akr1a1^{-/-}$ than wild-type mice ($P < 0.0001$) (Fig. 1e, f). The renal protection seen in $Akr1a1^{-/-}$ mice was lost in $Akr1a1^{-/-}Enos^{-/-}$ mice, indicating that protection by SCoR inhibition depends on NO. Conversely, $Enos^{-/-}$ mice were more susceptible to injury than wild-type mice, and deletion of AKR1A1 ($Akr1a1^{-/-}Enos^{-/-}$) counteracted this vulnerability (Fig. 1e, f, Extended Data Fig. 2d), indicating that protection by eNOS is associated with SNO-CoA. Tubular injury was attenuated in $Akr1a1^{-/-}$ mice compared with either $Akr1a1^{+/+}$ or $Akr1a1^{-/-}Enos^{-/-}$ mice (Fig. 1g, h, Extended Data Fig. 2e, f). As $Akr1a1^{-/-}$ mice have an ascorbate deficiency[7], chow diet was supplemented with 1% ascorbate, which normalized ascorbate levels, but had no effect on the AKI phenotype (Extended Data Fig. 3a–c). Collectively, our data support the idea that protection against AKI by eNOS-derived NO is associated with SNO-CoA bioactivity and governed by SCoR.

Knockout of SCoR improved survival following AKI (Fig. 1i). Female $Akr1a1^{-/-}$ mice exhibited the same protective phenotype as males, and both male and female $Akr1a1^{-/-}$ mice were also protected against lipopolysaccharide (LPS)-induced AKI (Extended Data Fig. 3d–i). We found that levels of endogenous SNOs (SNO-proteins) were significantly higher in injured kidneys of $Akr1a1^{-/-}$ than $Akr1a1^{+/+}$ mice ($P = 0.0221$) (Fig. 1j), whereas iron nitrosyl levels (a measure of NO production) were unchanged. These data suggest that protein *S*-nitrosylation by SNO-CoA protects against AKI.

Protein *S*-nitrosylation typically operates within multiprotein macro-complexes, in which SCoR may interact directly with SNO targets[4,8]. Most targets of SCoR in yeast are metabolic enzymes, and alterations in metabolism after AKI may have a protective role[5,9,10]. To identify protein targets of *S*-nitrosylation that mediate protection by the SNO-CoA–SCoR system, we combined three unbiased proteomic and metabolomic screening approaches. First, we coupled resin-assisted capture of SNO-proteins (SNO-RAC) with quantitative mass spectrometry. SNO-protein levels were higher in injured $Akr1a1^{-/-}$ kidneys than in $Akr1a1^{+/+}$ kidneys, and we found that 45 SNO-proteins were enriched by 1.4-fold or more (Fig. 2a, b, Supplementary Table 1). Second, we isolated the AKR1A1 interactome from mouse kidney extracts by immunoprecipitation, identifying 37 proteins (Supplementary Table 2). Notably, seven proteins overlapped with the nitrosoproteome (SNO-ome) identified by SNO-RAC, including the prominent metabolic enzyme pyruvate kinase M2 (PKM2) (Fig. 2c,

[1]Institute for Transformative Molecular Medicine, Department of Medicine, Case Western Reserve University and University Hospitals Cleveland Medical Center, Cleveland, OH, USA. [2]Center for Proteomics and Bioinformatics, Case Western Reserve University School of Medicine, Cleveland, OH, USA. [3]Division of Nephrology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [4]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [5]Division of Nephrology, Department of Medicine, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA. [6]Center for Vascular Biology Research, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, MA, USA. [7]Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA, USA. [8]Harrington Discovery Institute, University Hospitals Cleveland Medical Center, Cleveland, OH, USA. [9]These authors contributed equally: Rongli Zhang, Puneet Anand. *e-mail: jonathan.stamler@case.edu

© 2019 Springer Nature Limited. All rights reserved.

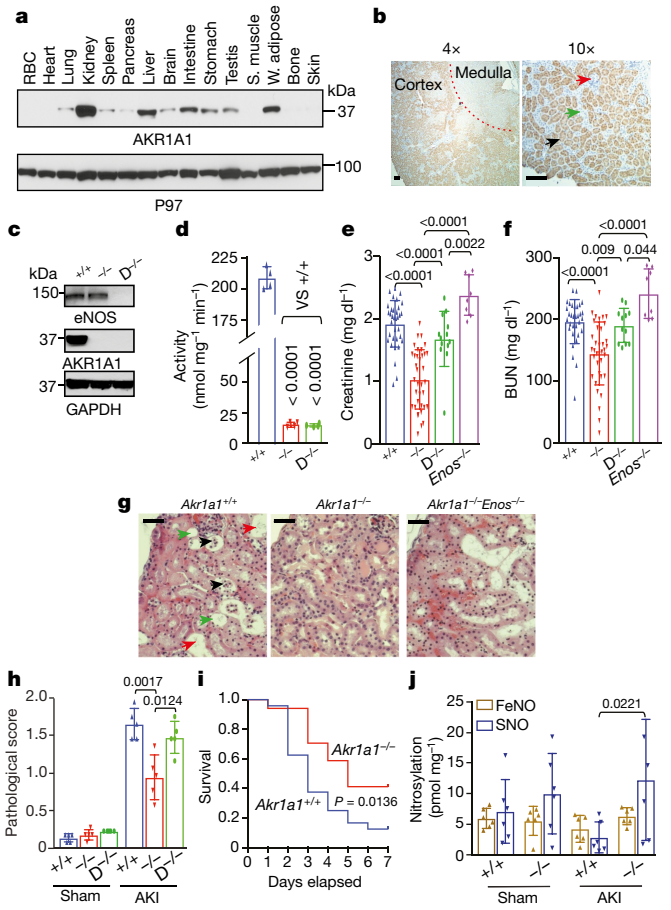


Fig. 1 | Knockout of AKR1A1 protects against AKI in a SNO-dependent manner. a, Expression of AKR1A1 in 15 mouse tissues. The AAA ATPase P97 is used as a loading control. S. muscle, skeletal muscle; W. adipose, white adipose tissue. **b**, Expression of AKR1A1 in proximal tubules. Immunostaining: 10× image derives from cortex area in 4× image. Black arrow, proximal tubule; green arrow, distal tubule; red arrow, glomerulus. Scale bars, 100 μm. **c**, Expression of AKR1A1 and eNOS in the kidneys of wild-type control ($^{+/+}$), *Akr1a1*$^{-/-}$ ($^{-/-}$) and *Akr1a1*$^{-/-}$*Enos*$^{-/-}$ ($D^{-/-}$) mice. **d**, NADPH-dependent SNO-CoA metabolizing activity ($n = 6$ mice per group). **e**, **f**, Serum creatinine and BUN after I/R-induced AKI (*Akr1a1*$^{+/+}$: 35 mice; *Akr1a1*$^{-/-}$: 36 mice; *Akr1a1*$^{-/-}$*Enos*$^{-/-}$: 13 mice; *Enos*$^{-/-}$: 8 mice). **g**, Haematoxylin and eosin stain of injured kidneys. AKI-induced renal tubular injury includes severe tubular lysis (black arrows), loss of brush borders (green arrows) and sloughed debris in the tubular lumen (red arrows). Scale bars, 50 μm. **h**, Pathological scores of tubular injury ($n = 5$ mice per group). **i**, Survival curve following I/R-induced AKI ($n = 24$ for *Akr1a1*$^{+/+}$ and 17 for *Akr1a1*$^{-/-}$; Gehan–Breslow–Wilcoxon test). **j**, Endogenous SNO-protein and iron nitrosyl (FeNO) levels (mean ± s.d.) in mouse kidneys ($n = 6$ mice per group). One-way ANOVA with Tukey post hoc test was used to detect significance in **e**, **f**, **h**, **j**. Numbers above square brackets show *P* values. Images in **a**–**c** and **g** are representative of two independently performed experiments with similar results. For gel source data, see Supplementary Fig. 1.

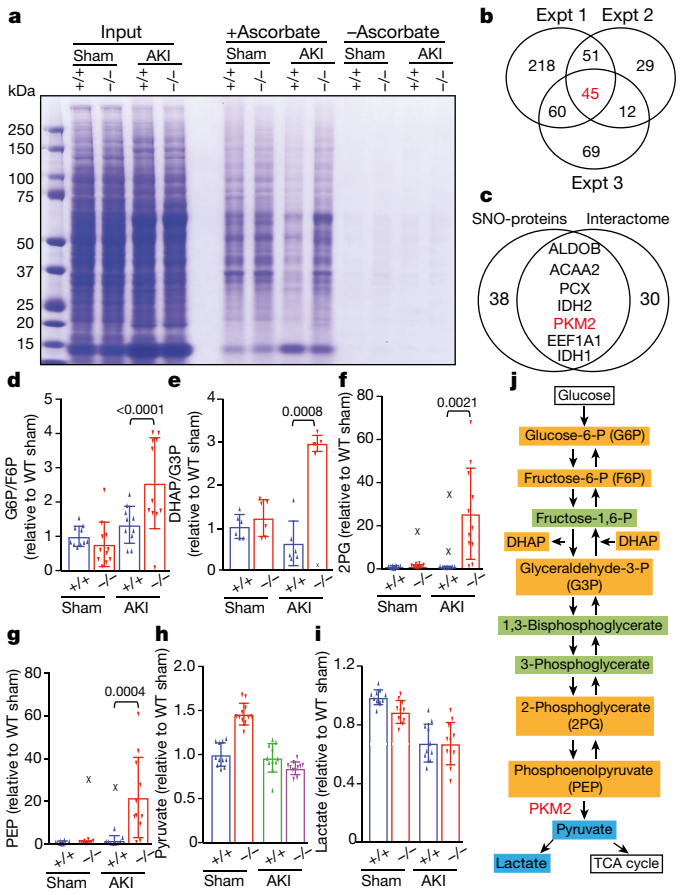


Fig. 2 | PKM2 is a major locus of regulation by the SNO-CoA–SCoR system. a, *S*-nitrosylated proteins (+Ascorbate) in the mouse kidneys. Image is representative of three independently performed experiments with similar results. –Ascorbate, control. **b**, *S*-nitrosylated proteins enriched more than 1.4-fold in injured kidneys from *Akr1a1*$^{-/-}$ versus injured kidneys from *Akr1a1*$^{+/+}$ mice in three independent experiments (Expt 1–3; SNO-RAC); injury induced by I/R. **c**, Proteins found in both the *S*-nitrosoproteome in **b** (left circle) and the AKR1A1 interactome (right circle). **d**–**i**, Mean ± s.d. glycolytic intermediates glucose-6-phosphate (G6P), fructose-6-phosphate (F6P), dihydroxyacetone phosphate (DHAP), glyceraldehyde-3-phosphate (G3P), 2-phosphoglycerate (2PG), phosphoenolpyruvate (PEP), pyruvate and lactate (G6P/F6P, 2PG, PEP, pyruvate and lactate: *Akr1a1*$^{+/+}$, $n = 10$ mice; *Akr1a1*$^{-/-}$, $n = 11$ mice; DHAP/ G3P: $n = 6$ mice per group). All data are normalized to wild-type (WT) sham; X, outliers. **j**, Glycolytic pathway metabolomics, comparing I/R-injured *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice. Intermediates highlighted in orange were increased, in blue were unchanged and in green were not identified. **d**–**i**, One-way ANOVA with Tukey post hoc test; numbers above square brackets show *P* values.

Supplementary Table 3). Third, we performed metabolic profiling following AKI (compared with sham injury) in *Akr1a1*$^{-/-}$ and *Akr1a1*$^{+/+}$ mice. Multiple upstream glycolytic intermediates accumulated in injured kidneys of *Akr1a1*$^{-/-}$ mice, whereas the downstream intermediates pyruvate and lactate did not accumulate (Fig. 2d–i). These data suggest a block at the last step in glycolysis—between phosphoenolpyruvate (PEP) and pyruvate—which is catalysed by PKM2 (Fig. 2j) (note that declines in pyruvate are likely to be prevented via multiple routes, including degradation of amino acids, conversion of lactate to pyruvate, and oxidative decarboxylation of L-malate[11,12]). Thus, PKM2 is identified as a SNO-CoA-regulated SNO-protein, a component of the AKR1A1 interactome, and a site of metabolic regulation by the SNO-CoA–SCoR system. These results point to inhibitory *S*-nitrosylation of PKM2 in injured kidneys of *Akr1a1*$^{-/-}$ mice.

To verify the regulation of PKM2 by SCoR, we measured the level and activity of *S*-nitrosylated PKM2 (SNO-PKM2) after AKI. SNO-PKM2 was higher in *Akr1a1*$^{-/-}$ than *Akr1a1*$^{+/+}$ kidneys, and increased SNO-PKM2 was associated with lower PKM2 activity; both increased SNO-PKM2 and decreased PKM2 activity were eNOS-dependent (Fig. 3a–c). Increased SNO-PKM2 and decreased PKM2 activity in *Akr1a1*$^{-/-}$ mice were also correlated with protection against endotoxin-induced AKI (Extended Data Fig. 3j–l). As further validation, we found that PKM2 interacted with AKR1A1 in HEK-293 cells, as it does in native kidneys, and that recombinant PKM2, but not other PK isoforms (PKM1 or PKLR), was directly inhibited by SNO-CoA (Fig. 3d, Extended Data Fig. 4a, b). Our data indicate that PKM2 activity after AKI is governed by SCoR-regulated *S*-nitrosylation.

PKM2 has 10 cysteine residues and individual mutation revealed that four cysteine residues—C152, C358, C423 and C424—accounted for

© 2019 Springer Nature Limited. All rights reserved.

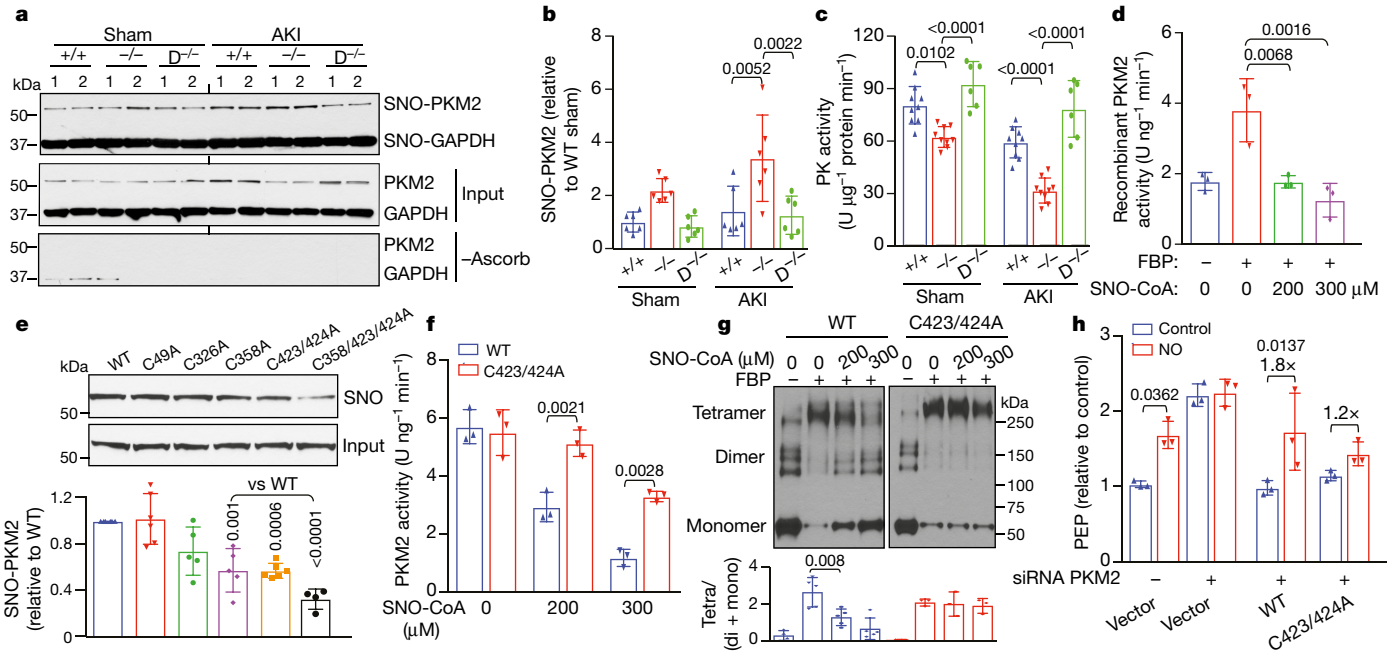


Fig. 3 | *S*-nitrosylation of renal PKM2 inhibits its activity by blocking tetramer formation. a, Endogenous *S*-nitrosylation of PKM2 (SNO-PKM2). SNO-GADPH and GAPDH (input) are used as internal controls. Image is representative of two independently performed experiments with similar results. **b**, Quantification of SNO-PKM2. SNO is normalized to PKM2 (input; $n = 6$ mice per group). **c**, Activity of endogenous pyruvate kinase (PK) ($n = 9$ mice per *Akr1a1*$^{+/+}$ or *Akr1a1*$^{-/-}$ group; $n = 6$ mice per *Akr1a1*$^{-/-}$*Enos*$^{-/-}$ (D$^{-/-}$) group. **d**, Activity of recombinant PKM2 proteins after SNO-CoA treatment ($n = 3$ independent experiments). FBP, fructose-1,6-biphosphate (PKM2 activator). **e**, SNO in PKM2 cysteine mutants expressed in HEK-293 cells ($n = 5$ independent experiments). **f**, Activity of recombinant PKM2(WT) and PKM2(C423/424A) after SNO-CoA treatment ($n = 3$ independent experiments). **g**, Tetramer/(dimer + monomer) distribution of recombinant PKM2(WT) and PKM2(C423/424A) after SNO-CoA treatment in vitro ($n = 5$ independent experiments for PKM2(WT); $n = 3$ independent experiments for PKM2(C423/424A)). **h**, Accumulation of glycolytic intermediate (PEP) in HEK-293 cells expressing Myc–PKM2(WT) or Myc–PKM2(C423/424A) after treatment with DETANO (NO) ($n = 3$ independent experiments). Results are presented as mean ± s.d. **b**–**h**, One-way ANOVA with Tukey post hoc test was used to detect significance; numbers above square brackets show *P* values.

measurable *S*-nitrosylation by eNOS (Fig. 3e, Extended Data Fig. 5a). PKM2 degradation was promoted by C152 mutation (Extended Data Fig. 5b–e). *S*-nitrosylation of PKM2 may therefore explain reduced PKM2 expression in *Akr1a1*$^{-/-}$ mice (Fig. 3a, Extended Data Fig. 4c, d). Oxidation of PKM2 at C358 can inhibit PKM2 activity[13,14]; however, C423 and C424 are newly discovered regulatory sites. Notably, C423 and C424 are encoded by the PKM2-specific alternative exon 10 and are localized at the interacting surface of the PKM2 tetramer (Extended Data Fig. 6a, b). The activity of PKM2(C423/424A) cannot be inhibited by SNO-CoA in vitro or by the NO donor DETA-NO in HEK-293 cells, confirming that cysteines 423 and 424 are the principal targets of NO (Fig. 3f, Extended Data Fig. 7a, b). Using purified proteins, we found that SNO-CoA inhibited the formation of tetrameric wild-type PKM2 but not tetrameric PKM2(C423/424A) (Fig. 3g). NO promoted the accumulation of PEP in cells expressing Myc-labelled wild-type PKM2 (Myc–PKM2(WT)) but not in those expressing Myc–PKM2(C423/424A) (Fig. 3h). Thus, *S*-nitrosylation of C423 and C424 is primarily responsible for inhibition of PKM2 by SNO-CoA.

We investigated how inhibition of a terminal step in glycolysis could confer protection against AKI. Multiple pentose phosphate pathway (PPP)-related intermediates were increased in *Akr1a1*$^{-/-}$ kidneys following AKI (Fig. 4a–d), and NO promoted higher accumulation of PPP intermediates in Myc–PKM2(WT) cells than in Myc–PKM2(C423/424A) cells (Extended Data Fig. 7d). PPP is a metabolic pathway for generating NADPH[15], which can increase glutathione (GSH) and activate anti-oxidant enzymes, lessening kidney injury[16], and we confirmed that the NADPH/NADP$^+$ ratio following AKI was higher in kidneys from *Akr1a1*$^{-/-}$ mice than from *Akr1a1*$^{+/+}$ or *Akr1a1*$^{-/-}$*Enos*$^{-/-}$ mice (Fig. 4e). Thus, inhibitory *S*-nitrosylation of PKM2 increases flux through the PPP.

Reactive oxygen species (ROS) are central mediators of AKI[17,18], and enhanced antioxidant defences can ameliorate AKI[19,20]. Tissue indicators of oxidative stress, oxidized GSH (GSSG)/GSH ratio and lipid peroxidation, were lower in injured kidneys from *Akr1a1*$^{-/-}$ mice than from *Akr1a1*$^{+/+}$ or *Akr1a1*$^{-/-}$*Enos*$^{-/-}$ mice (Fig. 4f, g) (without a change in total glutathione; Extended Data Fig. 7c). ROS levels may reflect mitochondrial dysfunction. However, levels of multiple tricarboxylic acid (TCA) cycle intermediates were similar in AKI-injured *Akr1a1*$^{-/-}$ and *Akr1a1*$^{+/+}$ mice, and ADP/ATP ratios were no different in Myc–PKM2(WT) than in Myc–PKM2(C423/424A) cells under NO treatment (Extended Data Fig. 8a–d). We conclude that inhibition of PKM2 by the SNO-CoA–SCoR system shunts metabolic intermediates through the PPP to alleviate oxidative stress and protect against AKI.

To establish conclusively the importance of PKM2 inhibition in protection against AKI and of metabolic reprogramming (PPP versus glycolytic flux) in renal protection, we generated mice in which PKM2 was knocked out specifically in renal tubular epithelial cells (*Pkm2*$^{-/-}$; Extended Data Fig. 9a). Levels of PKM2 were markedly reduced in kidneys of *Pkm2*$^{-/-}$ mice; however, levels of PKM1 were increased to compensate (Fig. 4h). Overall, pyruvate kinase activity in the kidney was reduced by about 40%, which recapitulates precisely PKM activity in the injured kidneys of *Akr1a1*$^{-/-}$ mice (Figs. 3c, 4i). Serum creatinine and BUN were significantly lower in *Pkm2*$^{-/-}$ than in wild-type mice following AKI ($P = 0.002$ for creatinine; $P = 0.0024$ for BUN) (Fig. 4j, k), indicating renal protection. Histological tubular injury was attenuated in *Pkm2*$^{-/-}$ compared with wild-type mice (Fig. 4l, m). Knockout of PKM2 improved survival (Extended Data Fig. 9b). NADPH/NADP$^+$ ratios and PEP levels, but not pyruvate levels, were increased in *Pkm2*$^{-/-}$ compared with wild-type mice (Fig. 4n, Extended Data Fig. 9c, d). The GSSG/GSH ratio and lipid peroxidation were lower in injured kidneys from *Pkm2*$^{-/-}$ mice than wild-type mice (Fig. 4o, p). These results confirm that inhibition of PKM2 shifts metabolic flux from energy-generating (glycolytic) to anti-oxidant (PPP) pathways to protect kidneys against AKI.

© 2019 Springer Nature Limited. All rights reserved.

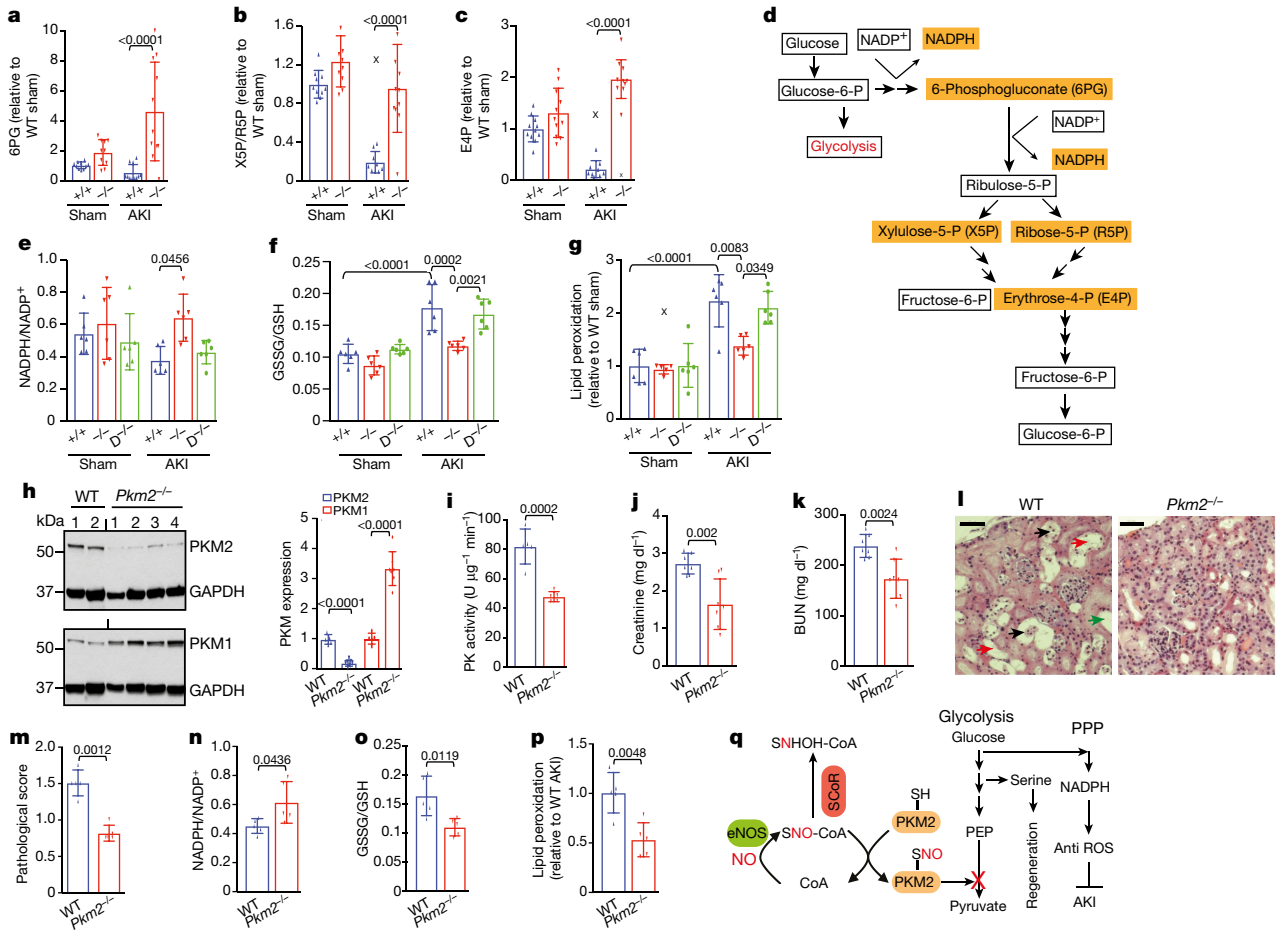


Fig. 4 | Inhibition of PKM2 by *S*-nitrosylation increases flux through the PPP and protects from AKI. a–c, Quantification of key PPP intermediates (6-phosphogluconate (6PG), xylulose-5-phosphate (X5P), ribose-5-phosphate (R5P), erythrose-4-phosphate (E4P); $n = 10$ mice per *Akr1a1*$^{+/+}$ group; $n = 11$ mice per *Akr1a1*$^{-/-}$ group). Injury induced with I/R. X, outliers. **d**, PPP; intermediates in orange were increased. **e**, **f**, NADPH/NADP$^+$ and GSSG/GSH ratios ($n = 6$ mice per group). **g**, Lipid peroxidation ($n = 6$ mice per group). **h**, Expression of PKM2 and PKM1. Quantification (right) is based on 5 mice per group. **i**, PK activity in kidneys ($n = 5$ mice per group). **j**, **k**, Serum creatinine and BUN after I/R-induced AKI ($n = 7$ mice per group). **l**, Haematoxylin and eosin staining of injured kidneys (arrows as in Fig. 1g). Image representative of two independently performed experiments with similar results. Scale bars, 50 μm. **m**, Pathological scores of tubular injury ($n = 5$ mice per group). **n–p**, NADPH/NADP$^+$, GSSG/GSH and lipid peroxidation ($n = 5$ mice per group). **q**, Working model showing how metabolic reprogramming by the SNO-CoA–SCoR system protects against kidney injury. After PKM2 inhibition (indicated by the red X), metabolites are shunted through serine and PPP pathways. **a–c**, **e–g**, One-way ANOVA with Tukey post hoc test; **h–k**, **m–p**, two-tailed Student's *t*-test; *P* values shown above square brackets. Results are presented as mean ± s.d.

The discovery of a physiological function for the SNO-CoA–SCoR system in mammals and in cellular metabolism in particular opens a new field of interdisciplinary inquiry. Notably, our results show that SNO-CoA has an essential role in metabolic regulation. SNO-CoA serves as an endogenous source of NO groups and thus as a mediator of protein *S*-nitrosylation, including of key metabolic enzymes (Supplementary Table 1). By coordinating metabolic flux through glycolysis versus PPP, the SNO-CoA–SCoR system regulates the balance between energy and reducing equivalents to protect against AKI (Fig. 4q). NO regulates glycolysis in neurons and glia, but the mechanism of this regulation has remained unclear[21–23]. Our findings suggest that SCoR-regulated, SNO-CoA-mediated protein *S*-nitrosylation may subserve metabolic signalling more broadly.

AKR1A1 has physiologically relevant SCoR activity in mammals, including dozens of SNO substrates. AKR1A1 has a role in ascorbic acid synthesis in rodents[7] and activity against γ-hydroxybutyric acid (GHB)-related aldehydes in vitro[24]. However, humans do not synthesize ascorbic acid, and AKR1A1 does not regulate GHB in vivo[25] (Extended Data Fig. 7g). Therefore, the primary function of AKR1A1 has been a mystery[26]. Our work indicates that the major function of AKR1A in mammals is to regulate NO-based metabolic signalling. Notably, eNOS-derived NO has previously been associated with both metabolic regulation and renal protection[1,2,27], but the molecular mechanisms of these effects were poorly understood. Our findings provide an unanticipated mechanistic basis for eNOS-derived NO protection in the kidney, mediated by SNO-CoA and governed by SCoR (Fig. 4q).

Pyruvate kinases catalyse the final step in glycolysis (Fig. 2j). PKM1 is expressed in organs with high energy requirements, including the heart, muscle and brain (Extended Data Fig. 6c), as a constitutively active tetramer, whereas PKM2 is expressed primarily in fetal (and tumour) cells, and can shift reversibly between a tetramer and a lower-activity dimer to reprogram metabolism for growth or survival[13,28,29]. Why PKM2 is expressed in some differentiated tissues and predominantly after AKI[9] (Extended Data Fig. 10a–c) has been unclear. We have shown that PKM2 expression enables protection by metabolic reprogramming, and coupled with reduced SCoR activity after AKI (Extended Data Fig. 2c), suggests physiological regulation. *S*-nitrosylation of PKM2 by SNO-CoA forces glucose flux into the PPP to detoxify ROS (Fig. 4q). PKM2 inhibition also increases synthesis of serine, a precursor for lipids, proteins and nucleotides[30], and serine levels are elevated in *Akr1a1*$^{-/-}$ mice following AKI (Extended Data Fig. 7e, f). Therefore, an additional advantage of metabolic programming via PKM2 may be in regenerating tissues after injury. In contrast to reversible regulation of PKM2 in AKI, irreversible PKM2

© 2019 Springer Nature Limited. All rights reserved.

inactivation is associated with diabetic nephropathy[14]. Thus, inhibition of SCoR and/or PKM2 may be advantageous therapeutically in acute injurious conditions, including AKI.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0749-z.

Received: 20 April 2018; Accepted: 31 October 2018;
Published online 28 November 2018.

1. Wang, W. et al. Endothelial nitric oxide synthase-deficient mice exhibit increased susceptibility to endotoxin-induced acute renal failure. *Am. J. Physiol. Renal Physiol.* **287**, F1044–F1048 (2004).
2. Forbes, M. S., Thornhill, B. A., Park, M. H. & Chevalier, R. L. Lack of endothelial nitric-oxide synthase leads to progressive focal renal injury. *Am. J. Pathol.* **170**, 87–99 (2007).
3. Seth, D. et al. A multiplex enzymatic machinery for cellular protein *S*-nitrosylation. *Mol. Cell* **69**, 451–464 (2018).
4. Stomberski, C. T., Hess, D. T. & Stamler, J. S. Protein *S*-nitrosylation: determinants of specificity and enzymatic regulation of *S*-nitrosothiol-based signaling. *Antioxid. Redox Signal*. https://doi.org/10.1089/ars.2017.7403 (2018).
5. Anand, P. et al. Identification of *S*-nitroso-CoA reductases that regulate protein *S*-nitrosylation. *Proc. Natl Acad. Sci. USA* **111**, 18572–18577 (2014).
6. Nagasu, H. et al. Endothelial dysfunction promotes the transition from compensatory renal hypertrophy to kidney injury after unilateral nephrectomy in mice. *Am. J. Physiol. Renal Physiol.* **302**, F1402–F1408 (2012).
7. Gabbay, K. H. et al. Ascorbate synthesis pathway: dual role of ascorbate in bone homeostasis. *J. Biol. Chem.* **285**, 19510–19520 (2010).
8. Jia, J. et al. Target-selective protein *S*-nitrosylation by sequence motif recognition. *Cell* **159**, 623–634 (2014).
9. Lan, R. et al. Mitochondrial pathology and glycolytic shift during proximal tubule atrophy after ischemic AKI. *J. Am. Soc. Nephrol.* **27**, 3356–3367 (2016).
10. Kang, H. M. et al. Defective fatty acid oxidation in renal tubular epithelial cells has a key role in kidney fibrosis development. *Nat. Med.* **21**, 37–46 (2015).
11. Chang, G. G. & Tong, L. Structure and function of malic enzymes, a new class of oxidative decarboxylases. *Biochemistry* **42**, 12721–12733 (2003).
12. Gray, L. R., Tompkins, S. C. & Taylor, E. B. Regulation of pyruvate metabolism and human disease. *Cell. Mol. Life Sci.* **71**, 2577–2604 (2014).
13. Anastasiou, D. et al. Inhibition of pyruvate kinase M2 by reactive oxygen species contributes to cellular antioxidant responses. *Science* **334**, 1278–1283 (2011).
14. Qi, W. et al. Pyruvate kinase M2 activation may protect against the progression of diabetic glomerular pathology and mitochondrial dysfunction. *Nat. Med.* **23**, 753–762 (2017).
15. Fan, J. et al. Quantitative flux analysis reveals folate-dependent NADPH production. *Nature* **510**, 298–302 (2014).
16. Kim, J. et al. Role of cytosolic $NADP^+$-dependent isocitrate dehydrogenase in ischemia-reperfusion injury in mouse kidney. *Am. J. Physiol. Renal Physiol.* **296**, F622–F633 (2009).
17. Sharfuddin, A. A. & Molitoris, B. A. Pathophysiology of ischemic acute kidney injury. *Nat. Rev. Nephrol.* **7**, 189–200 (2011).
18. Granger, D. N. & Kvietys, P. R. Reperfusion injury and reactive oxygen species: the evolution of a concept. *Redox Biol.* **6**, 524–551 (2015).
19. Zuk, A. & Bonventre, J. V. Acute kidney injury. *Annu. Rev. Med.* **67**, 293–307 (2016).
20. Ratliff, B. B., Abdulmahdi, W., Pawar, R. & Wolin, M. S. Oxidant mechanisms in renal injury and disease. *Antioxid. Redox Signal.* **25**, 119–146 (2016).
21. Benavides, G. A., Liang, Q., Dodson, M., Darley-Usmar, V. & Zhang, J. Inhibition of autophagy and glycolysis by nitric oxide during hypoxia-reoxygenation impairs cellular bioenergetics and promotes cell death in primary neurons. *Free Radic. Biol. Med.* **65**, 1215–1228 (2013).
22. Almeida, A., Moncada, S. & Bolaños, J. P. Nitric oxide switches on glycolysis through the AMP protein kinase and 6-phosphofructo-2-kinase pathway. *Nat. Cell Biol.* **6**, 45–51 (2004).
23. Bolaños, J. P., Delgado-Esteban, M., Herrero-Mendez, A., Fernandez-Fernandez, S. & Almeida, A. Regulation of glycolysis and pentose-phosphate pathway by nitric oxide: impact on neuronal survival. *Biochim. Biophys. Acta* **1777**, 789–793 (2008).
24. O'Connor, T., Ireland, L. S., Harrison, D. J. & Hayes, J. D. Major differences exist in the function and tissue-specific expression of human aflatoxin B1 aldehyde reductase and the principal human aldo-keto reductase AKR1 family members. *Biochem. J.* **343**, 487–504 (1999).
25. Alzeer, S. & Ellis, E. M. The role of aldehyde reductase AKR1A1 in the metabolism of γ-hydroxybutyrate in 1321N1 human astrocytoma cells. *Chem. Biol. Interact.* **191**, 303–307 (2011).
26. Barski, O. A., Tipparaju, S. M. & Bhatnagar, A. The aldo-keto reductase superfamily and its role in drug metabolism and detoxification. *Drug Metab. Rev.* **40**, 553–624 (2008).
27. Doulias, P. T., Tenopoulou, M., Greene, J. L., Raju, K. & Ischiropoulos, H. Nitric oxide regulates mitochondrial fatty acid metabolism through reversible protein *S*-nitrosylation. *Sci. Signal.* **6**, rs1 (2013).
28. Anastasiou, D. et al. Pyruvate kinase M2 activators promote tetramer formation and suppress tumorigenesis. *Nat. Chem. Biol.* **8**, 839–847 (2012).
29. Christofk, H. R. et al. The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature* **452**, 230–233 (2008).
30. Ye, J. et al. Pyruvate kinase M2 promotes de novo serine synthesis to sustain mTORC1 activity and cell proliferation. *Proc. Natl Acad. Sci. USA* **109**, 6904–6909 (2012).

Acknowledgements We thank H. Fujioka for mitochondrial analyses, J. Mikulan and A. Kresak for histology and immunostaining, C. Geng for assistance with statistics, and J. Reynolds, D. Hess, R. Premont and D. Seth for discussions. This work is supported by National Institutes of Health grants DK119506, HL075443, HL128192 and HL126900.

Reviewer information *Nature* thanks H. Christofk, C. Lowenstein and G. Remuzzi for their contribution to the peer review of this work.

Author contributions H.-L.Z., P.A. and J.S.S. designed the study. H.-L.Z. carried out most of the experiments and analysed the results. R.Z. performed AKI surgery. P.A. prepared samples for iTRAQ LC–MS/MS and metabolomics. C.T.S. prepared samples for the photolysis-chemiluminescence assay. A.H. and P.A. purified SCoR from bovine kidney. Z.Q. handled mice. L.W. performed quantitative iTRAQ LC–MS/MS. E.P.R. and S.M.P. contributed to project conception and carried out metabolomics analyses. S.A.K. contributed to project conception and performed histological stains. H.-L.Z. and J.S.S. wrote the manuscript with input from all authors.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0749-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0749-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to J.S.S.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

Mice. Mouse studies were approved by the Case Western Reserve University Institutional Care and Use Committee (IACUC). Housing and procedures complied with the Guide for the Care and Use of Laboratory Animals and the American Veterinary Medical Association guidelines on euthanasia. *Akr1a1*$^{+/-}$ mice were made by Deltagen. In brief, to knock out AKR1A1 (SCoR) in embryonic stem cells, the *Akr1a1*$^{+/-}$ allele was first created by insertion of a LacZ-Neo cassette in place of exon 2 of the *Akr1a1* gene, disrupting in-frame translation of AKR1A1 (Extended Data Fig. 1g, h). F1 mice were generated by breeding chimeric male mice with C57BL/6 females. F2 homozygous mutant mice were produced by intercrossing F1 heterozygous males and females. Wild-type littermates produced by crossing *Akr1a1*$^{+/-}$ and *Akr1a1*$^{+/-}$ mice were used as breeding pairs to generate control mice (*Akr1a1*$^{+/+}$). To generate AKR1A1 and eNOS double knockout (*Akr1a1*$^{-/-}$*Enos*$^{-/-}$) mice, male *Akr1a1*$^{-/-}$ mice were crossed with female *Enos*$^{-/-}$ mice, obtained from Jackson Laboratory. To generate renal tubular epithelial cell-specific PKM2-knockout mice (*Pkm2*$^{fl/fl}$; *KSP-cre* or *Pkm2*$^{-/-}$), conditional PKM2-knockout mice (*Pkm2*$^{fl/fl}$) were crossed with *KSP-cre* mice (both obtained from Jackson Laboratory). Wild-type littermates (*Pkm2*$^{+/+}$;*KSP-cre*) produced by intercrossing *Pkm2*$^{fl/+}$;*KSP-cre* parents were used as breeding pairs to generate control mice. *Pkm2*$^{fl/fl}$ mice possess *loxP* sites flanking exon 10 of the PKM gene, which when deleted forces PKM transcripts to splice as PKM1[31]. In *KSP-cre* mice, the cadherin 16 promoter drives Cre to specifically express in epithelial cells of renal tubules[32]. Genotyping of *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice was performed according to the PCR protocol from Deltagen using the following primers: common forward: 5′-GCAGAGATTCAACAAGTCTCCCCTC-3′; mutant reverse: 5′-GGGCCAGCTCATTCCTCCCACTCAT-3′; common reverse: 5′-AGCTAAGGCTCCGAGCAGTGCTAAC-3′. Genotyping of *Akr1a1*$^{-/-}$*Enos*$^{-/-}$ mice and *Pkm2*$^{-/-}$ mice was performed according to the PCR protocol from Jackson laboratory using the following primers: *Enos*$^{-/-}$ mutant forward: 5′-AATTCGCCAATGACAAGACG-3′; *Enos*$^{-/-}$ wild-type forward: 5′-AGGGGAACAAGCCCAGTAGT-3′; *Enos*$^{-/-}$ common reverse: 5′-CTTGTCCCC TAGGCACCTCT-3′; *Pkm2*$^{fl/fl}$ forward: 5′-CCTTCAGGAAGACAGCCAAG-3′; *Pkm2*$^{fl/fl}$ reverse: 5′-AGTGCTGCCT GGAATCCTCT-3′; *KSP-cre* forward: 5′-GCAGATCTGGCTCTCCAAAG-3′; *KSP-cre* reverse: 5′-AGGCAAATTT TGGTGTACGG-3′.

Acute kidney injury. AKI surgery was carried out as described previously[33]. Mice of similar age (9–11 weeks) and body weight (male: 25–28 g; female: 22–25 g) were used for surgery. The mice were anaesthetized with isoflurane (1–3%) in oxygen and then anaesthesia was maintained by adjusting isoflurane (0.75–2.0%) as needed. The fur in the surgical area was removed with clippers and the skin sterilized with 3 times alternating washes of betadine (or chlorhexidine) and alcohol. The mouse was placed on a thermostatic station during surgery. The skin and muscle were cut open along the back to expose both right and left kidneys. Gentle blunt dissection revealed the kidney and a Q-tip was used to mobilize and exteriorize the kidney. A 6-0 silk suture was used to clamp the pedicle to block the blood flow to the kidney to induce renal ischaemia for 23 min in male mice or 50 min in female mice, then the sutures were released to allow reperfusion. The identical steps were performed on both kidneys. A silk suture was used to close the muscle layer of the incision followed by the closure of the skin wound with vicryl. Immediately after the wound closure, 20 ml/kg sterile saline was given intraperitoneally to each mouse. The animal was then kept on a heating pad until it gained full consciousness before being returned to home cage. Mice subjected to surgery without clamping the pedicle were used as sham controls. Mortality at 24 h after AKI for WT, *Akr1a1*$^{-/-}$, *Akr1a1*$^{-/-}$*Enos*$^{-/-}$ and *Enos*$^{-/-}$ mice is shown in Extended Data Fig. 2g. Serum creatinine and BUN were determined after 24 h of reperfusion upon removal of the kidney (when larger volumes of blood can be collected). Serum creatinine and BUN were measured at University Hospital's Clinical Laboratories.

For the LPS-induced AKI model, LPS (O111:B4, Sigma) in saline (0.9%) was injected intraperitoneally into each mouse (10 mg/kg). Immediately after the injection of LPS, 20 ml/kg sterile saline was given intraperitoneally to each mouse. Serum creatinine and BUN were determined after 16 h.

SNO-RAC. SNO-RAC was carried out as described previously[34]. Mouse kidneys were mechanically homogenized in lysis buffer (1 mg/5 μl lysis buffer) containing 100 mM HEPES/1 mM EDTA/100 μM neocuproine (HEN), 50 mM NaCl, 0.1% (vol/vol) Nonidet P-40 (NP-40), the thiol-blocking agent 0.2% *S*-methylmethanethiosulfonate (MMTS), 1 mM PMSF and protease inhibitors (Roche). After centrifugation (20,000*g*, 4 °C, 20 min, × 2), SDS and MMTS were added to the supernatants to 2.5% and 0.2% respectively, and incubated at 50 °C for 20 min. Proteins were precipitated with −20 °C acetone, and re-dissolved in 1 ml HEN/1% SDS. Precipitation of proteins was repeated with −20 °C acetone, and the final pellets were resuspended in HEN/1% SDS and protein concentrations determined using the bicinchoninic acid (BCA) method. Total lysates (2 mg) were incubated with freshly prepared 50 mM ascorbate and 50 μl thiopropyl-sepharose (50% slurry) and rotated end-over-end in the dark for 4 h. The bound SNO proteins were sequentially washed with HEN/1% SDS and 10% HEN/0.1% SDS; SNO proteins were then eluted with 10% HEN/1% SDS/10% β-meracaptoethanol and analysed by SDS–PAGE and immunoblotting.

iTRAQ-coupled SNO-RAC. iTRAQ (isobaric tags for relative and absolute quantitation)-coupled SNO-RAC was carried out as described previously[5]. Extracts of the kidneys were prepared, and SNO-RAC (4 mg of protein per sample) was carried out as described above. SDS–PAGE gels were Coomassie-blue stained, and lanes were separated into eight segments top-to-bottom and collected in two 1.5-ml tubes. Then, 500 μl of 50% acetonitrile (ACN)/50% 100 mM ammonium bicarbonate was used to wash gel bands for more than 5 h while vortexing. After removal of washing buffer, 400 μl of 100% acetonitrile was added to gel pieces and vortexed for 10 min. After removal of ACN, gel pieces were dried in a speed vacuum dryer for 10 min. Subsequently, 200 μl of 10 mM dithiothreitol (DTT) was added to dry gel pieces and vortexed for 45 min. After removal of the DTT buffer, 200 μl of 55 mM iodoacetamide (IAA) was added to the gel pieces, followed by incubating for 45 min in the dark. After removal of IAA buffer, 400 μl of 1× iTRAQ dissolution solution and 400 μl ACN were used to wash the gel pieces alternately twice. Gel pieces were dried for 10 min in a speed vacuum dryer. Next, 500 ng trypsin in 150 μl 1× iTRAQ buffer was added to dried gel pieces on ice for 30 min, and then incubated overnight at 37 °C. Supernatant from the digested protein solution was transferred to a 1.5-ml tube using gel-loading tips. Then, 200 μl extraction buffer of 60% ACN/5% formic acid was added to gel pieces, vortexed for 30 min, and sonicated for 15 min. The supernatant containing peptide extracts was transferred to a 1.5-ml tube, and extractions were repeated two more times. The final digested solution was dried completely. iTRAQ labelling was performed according to the instructions of iTRAQ Reagents, 4plex Applications Kit. In brief, 30 μl of iTRAQ dissolution buffer (10×) was added to each sample tube (pH > 7), and then iTraq labelling reagents (114, 115, 116, 117) were added to separate sample tubes: one reagent to one sample tube. Labelling reactions were vortexed for more than 5 h at room temperature to ensure complete labelling efficiency. The four labelled samples were mixed together and dried. Next, 160 μl of 5% ACN containing 0.5% TFA was added to the mixed labelled sample and cleaned using C18 ziptips. In brief, C18 tips were wetted 5 times with 20 μl of 50% ACN each time, equilibrated with 100 μl of 5% ACN containing 0.5% TFA. Samples were then loaded to the tip by drawing and expelling for 50 cycles to ensure complete binding. The tips were then washed with 20 μl of 5% ACN containing 0.5% TFA 10 times. Peptides were eluted from tips with 20 μl of 60% ACN containing 0.1% formic acid three times, and eluates were combined and dried for liquid chromatography with tandem mass spectrometry (LC–MS/MS) analysis.

Immunoprecipitation. In brief, 15 μg of AKR1A1 polyclonal antibody (Proteintech) was incubated with 50 μl Protein G Sepharose (GE) (1:1 slurry) at 4 °C overnight. After washing with NETN buffer (150 mM NaCl, 20 mM Tris-HCl (pH 8.0), 0.5 mM EDTA, 0.5% (v/v) NP-40, 1 mM PMSF and protease inhibitors cocktail)) three times, AKR1A1 antibody bound to Protein G Sepharose was ready for immunoprecipitation. Mouse kidneys were mechanically homogenized in EBC lysis buffer (120 mM NaCl, 20 mM Tris-HCl (pH 8.0), 0.5 mM EDTA, 0.5% (v/v) NP-40, 1 mM PMSF and protease inhibitors cocktail (1 mg tissue/5 μl lysis buffer)). After centrifugation (20,000*g*, 4 °C, 20 min, × 2), 2 ml (20 mg/ml) supernatant was pre-cleared by incubation with 50 μl Protein G Sepharose (1:1 slurry) for 1 h at 4 °C. After being spun down at 1,000*g* for 1 min, the supernatant was transferred into new tubes and incubated with 50 μl anti-AKR1A1 antibody-Protein G Sepharose (1:1 slurry) for 5 h at 4 °C. Beads were washed with NETN buffer and proteins were eluted with 50 μl 0.1 M glycine (pH 2.5) for 10 min at room temperature with shaking. Following centrifugation at 1,000*g* for 2 min, the eluate was neutralized by the addition of 5 μl Tris-HCl (1.0 M), pH 8.0. Proteins were identified by LC–MS/MS analysis. Co-immunoprecipitation (co-IP) was carried out in HEK-293 cells overexpressing V5–AKR1A1 and Myc–PKM2, by co-transfection using Lipofectamine 2000. Cells were collected and lysed in EBC lysis buffer. Anti-Myc affinity gel (Sigma) was used for co-IP.

LC–MS/MS analysis. Digested peptides were separated using a UPLC (Waters) with a Nano-ACQUITY UPLC BEH300 C18 column. Separated peptides were continuously injected into an Orbitrap Elite hybrid mass spectrometer (Thermo Finnigan) using a nanospray emitter (10 μm, New Objective). A linear gradient using mobile phase A (0.1% formic acid in water) and B (100% acetonitrile) was used at a flow rate of 0.3 μl/min, starting with 1% mobile phase B and increasing to 40% B at 65 min for protein interaction identification, or increasing to 40% B at 130 min for iTRAQ experiments. All mass spectrometry data were acquired in positive ion mode. For protein interaction identification, a full MS scan (*m*/*z* 350–1,800) at a resolution of 120,000 was conducted, and twenty MS2 scans (*m*/*z* 350–1,800) were selected from the twenty most intense peptide peaks of full MS scans. Collision-induced dissociation (CID) cleavage mode was performed at a normalized collision energy of 35%. For iTRAQ experiments, a full MS scan (*m*/*z* 300–1,800) at resolution of 120,000 was conducted, and ten MS2 scans

© 2019 Springer Nature Limited. All rights reserved.

(m/z 100–1,600) were activated from the five most intense peptide peaks of the full MS scans. CID and higher-energy collisional dissociation (HCD) cleavage modes were performed alternately on the same peptides selected from full MS scans. The MS2 resolution of HCD was 15,000. The bioinformatic software MassMatrix was used to search MS data against a database composed of sequences of mouse proteins from Uniprot and their reversed sequences as a decoy database. Modifications including oxidation of methionine and labelling of cysteine (IA modifications) were selected as variable modifications in searching. For iTRAQ labelling searching, the MS tags of N terminus, Lys and/or Tyr were selected as variable modifications to test labelling efficiency and fixed modification for iTRAQ analysis. Trypsin was selected as the in silico enzyme to cleave proteins after Lys and Arg. Precursor ion searching was within 10 p.p.m. mass accuracy and product ions within 0.8 Da for CID cleavage mode and 0.02 Da for HCD cleavage mode. 95% confidence interval was required for protein identification.

Cloning, expression, and purification of recombinant PKM2. The mammalian cell expression plasmid pCMV-PKM2 (Human) was obtained from Origene. The mammalian cell expression plasmid pcDNA-AKR1A1 was constructed by PCR cloning of human AKR1A1. pCMV-PKM2 cysteine mutants were generated using a QuikChange II Site-Directed Mutagenesis Kit (Agilent). For purification of recombinant PKM2, cDNA encoding PKM2(WT) or PKM2(C423A/424A) was cloned into pET21b (Novagen) to introduce a C-terminal 6×His tag on the expressed protein. The recombinant PKM2 proteins were purified from BL21-CodonPlus Competent *Escherichia coli* cells (Agilent). Overnight *E. coli* cultures were sub-cultured into 1 l of LB medium at 5%. At an $OD_{600\ nm}$ of 0.5, cultures were induced with 100 mM IPTG and grown for a further 4 h at 28 °C. Cultures were centrifuged at 4,000*g* for 10 min to harvest the cells. Cell pellets from 1 l cultures were lysed in 10 ml of 1×PBS buffer containing 1 mM PMSF and protease-inhibitor cocktail by sonication. After centrifugation at 14,500*g* for 20 min, the supernatant was collected. The lysate was diluted in 30 ml 1×PBS buffer containing 1 mM PMSF and protease-inhibitor cocktail and incubated with 1 ml of Ni-NTA agarose at 4 °C for 1 h with rotation. The slurry was then poured into empty PD-10 columns (GE Healthcare). The beads were washed with 100 ml of 50 mM NaH_2PO_4, 300 mM NaCl buffer containing 20 mM imidazole. Elution was done in 2 ml of 50 mM NaH_2PO_4, 300 mM NaCl buffer with 250 mM imidazole. Buffer was exchanged with modified Roeder D ((20 mM HEPES (pH 7.9), 20% (v/v) glycerol, 0.1 M KCl, 0.2 mM EDTA)) through a Microcon centrifugal filter device (Millipore).

Cell culture, siRNA and related treatments. HEK-293 cells were obtained from ATCC (Manassas, Virginia) and tested for mycoplasma contamination. HEK-293 transfection was done as described previously[35]. siRNA-mediated protein depletion was performed in HEK-293 cells. Two custom PKM2 siRNAs that target the 3′ UTR of human PKM2 (5′-CCAGAUGGCAAGAGGGUG-3′ and 5′-GAUCAACGCCUCACUGAAA-3′) were obtained from Dharmacon. siRNA oligonucleotides (60 pmol per 10 cm plate) were transfected into HEK-293 cells using Lipofectamine RNAiMAX (Invitrogen) according to the manufacturer's protocol. After 24 h, 5 μg pCMV-control, pCMV-PKM2-WT or pCMV-PKM2-C423/424A were co-transfected with siRNA oligonucleotides (60 pmol per 10 cm plate) into HEK-293 cells using Lipofectamine 2000. After 24 h, cells were treated with 500 μM DETA-NO for 20 h. Cells were then collected for assay.

Pyruvate, GHB, PEP, 6PG, ATP, ADP and serine measurement. The amount of pyruvate was measured using a Pyruvate Assay Kit (Sigma). Kidneys from *Akr1a1*$^{+/+}$, *Akr1a1*$^{-/-}$, *Pkm2*$^{+/+}$ or *Pkm2*$^{-/-}$ mice (sham operation or AKI) were mechanically homogenized in pyruvate assay buffer (1 mg/5 μl buffer). After extracts were clarified by centrifugation (20,000*g*, 4 °C, 20 min, ×2), supernatant was used for assay. GHB in the serum of *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice was measured using the GHB enzymatic assay kit from Buhlmann. To measure PEP, 6PG, ATP, ADP and serine in HEK-293 cells, 1×10^6 cells were lysed in corresponding buffer. The amount of PEP, 6PG, ATP, ADP and serine were measured using PEP Colorimetric/Fluorometric Assay Kit (Sigma), 6 Phosphogluconate Assay kit (abcam), ATP Colorimetric/Fluorometric Assay Kit (Sigma), ADP Colorimetric/Fluorometric Assay Kit (Sigma) and DL-Serine Assay kit (Fluorometric) (Biovision), respectively.

Assay of NADPH-dependent SNO-CoA reductase activity in mouse. Kidneys from *Akr1a1*$^{+/+}$, *Akr1a1*$^{-/-}$ and *Akr1a1*$^{-/-}$*Enos*$^{-/-}$ mice were mechanically homogenized in lysis buffer (50 mM phosphate buffer, pH 7.0, 150 mM NaCl, 0.1 mM EDTA, 0.1 mM DTPA, 1 mM PMSF, and protease inhibitor mixture (Roche)). Extracts were clarified by centrifugation (20,000*g*, 4 °C, 20 min, ×2), and protein concentration was determined by BCA assay. The NADPH-dependent SNO-CoA reductase activity was determined spectrophotometrically as described previously[5]. In brief, the assays were performed in 50 mM phosphate buffer (pH 7.0; containing 0.1 mM EDTA and DTPA) with 0.2 mM SNO-CoA and 0.1 mM NADPH. Reactions were initiated by the addition of lysate and allowed to proceed for 1 min. All assays were performed in triplicate.

Photolysis/chemiluminescence. Kidneys from *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice were mechanically homogenized in lysis buffer (50 mM phosphate buffer, pH 7.0, 150 mM NaCl, 0.1 mM EDTA, 0.1 mM DTPA, 1 mM PMSF, and protease inhibitor mixture (Roche)). Extracts were clarified by centrifugation (20,000*g*, 4 °C, 20 min, ×2), and protein concentration was determined by BCA assay. Measurements of XNO/SNO (where XNO is predominantly metal-NO (MNO)) in lysates were done using photolysis/chemiluminescence essentially as described[36]. In brief, NO released from MNO/SNO by UV-photolysis was detected by chemiluminescence generated by the reaction of NO with ozone. Pre-treatment of samples with $HgCl_2$ (1 mM) (Hg^{2+}-coupled photolysis/chemiluminescence) removes SNO specifically and allows differentiation between SNO and other photolysable NO species (predominantly MNO).

Histological analysis. Kidney samples were fixed with 4% PFA for 24 h, dehydrated and embedded into paraffin blocks. Formalin-fixed, paraffin-embedded blocks were sectioned and stained with haematoxylin and eosin. For immunohistochemistry staining, paraffin sections were dewaxed and rehydrated. Antigen retrieval was performed by boiling sections in 0.01 M sodium citrate buffer (pH 6.0) for 20 min, then sections were washed three times with PBS. Rabbit polyclonal anti-AKR1A1 (15054-1-AP, Proteintech Group, 1:100) or rabbit polyclonal anti-PKM2 (ABS245, EMD Millipore, 1:100) was dropped onto sections and incubated at 4 °C overnight. After washing with PBS, secondary antibody (HRP-associated goat anti-rabbit) was dropped onto sections and incubated at room temperature for 1 h. Diaminobenzidine (DAB) was used for coloration. More than ten microscopic fields obtained from each animal were selected for quantitative analysis. Renal histopathologic alterations were evaluated and graded on a 0 to 2 scale as described previously[37,38].

Electron microscopy. Mice were perfused transcardially with quarter-strength Karnovsky's fixative solution at a flow rate of 10 ml/min for 10 min. Small pieces of kidney were immersed in triple aldehyde-DMSO[39]. After rinsing in 0.1 M phosphate buffer (pH 7.3), they were post-fixed in ferrocyanide-reduced osmium tetroxide. Another water rinse was followed by an overnight soak in acidified uranyl acetate. After again rinsing in distilled water, the tissue blocks were dehydrated in ascending concentrations of ethanol, passed through propylene oxide, and embedded in Poly/Bed resin. Thin sections were sequentially stained with acidified uranyl acetate followed by a modification of Sato's triple lead stain[40]. These sections were examined in a FEI Tecnai Spirit (T12) transmission electron microscope with a Gatan US4000 4k×4k CCD.

PK activity. PKM activity was measured on the basis of generation of pyruvate, which was oxidized by pyruvate oxidase to produce colour ($\lambda = 570$ nm). To measure PKM1, PKLR, PKM2(WT) and PKM2(C423/424A) (where both Cys are replaced by Ala) protein activity in vitro, DTT in PKM1 (Sigma) and PKLR (R&D systems) buffer was removed using Amicon Ultra-0.5 ml Centrifugal Filters. In brief, 250 ng recombinant PKM1, PKLR, PKM2(WT) or PKM2(C423/424A) was pre-incubated with substrate 2 μl fructose-1,6-bisphosphate (FBP) (250 μM) in 2 ml dialysis buffer (20 mM Tris-HCl (pH 7.9), 20% (v/v) glycerol, 0.1 M KCl, 0.2 mM EDTA), followed by dialysis to remove the free FBP in 2 l dialysis buffer. PKM2–FBP complex or PKM2(C423/424A)–FBP complex was treated with 200–300 μM SNO-CoA for 10 min at room temperature, after which the activity of PKM2(WT) and PKM2(C423/424A) was measured. To measure PKM1 and PKLR activity, 10 ng PKM1–FBP complex, PKLR–FBP complex or PKM2–FBP complex was treated with 0, 50, 100, 200 or 300 μM SNO-CoA for 10 min at room temperature before assay. To measure PK activity in kidney, kidneys were mechanically homogenized in lysis buffer (50 mM phosphate buffer, pH 7.0, 150 mM NaCl, 0.1 mM EDTA, 0.1 mM DTPA, 1 mM PMSF, and protease inhibitor mixture (Roche)). Extracts were clarified by centrifugation (20,000*g*, 4 °C, 20 min, ×2), and protein concentration was determined by BCA assay. Then, 10 μl (0.1 μg/μl) lysate was used to measure PKM2 activity. Assay of PKM2 activity followed the protocol of the Pyruvate Kinase Activity Colorimetric/Fluorometric Assay Kit (Biovision).

PKM2 dimer and tetramer formation. The assay of PKM2 dimer and tetramer in vitro was carried out as previously described[41]. In brief, after 40 ng PKM2–FBP complex was treated with 200–300 μM SNO-CoA for 10 min at room temperature, 5 μl fresh glutaraldehyde (50%) was added to a reaction mixture containing 100 mM HEPES (pH 7.5) for 5 min at 37 °C. The cross-linking reaction was terminated by addition of 5 μl 1M Tris-HCl (PH 8.0). Equal amounts of protein were separated using a 4–20% Criterion Precast Midi Protein Gel (BIO-RAD) and monomer, dimer and tetramer forms of PKM2 were detected with PKM2 antibody (sc-365684, Santa Cruz Biotechnology).

Western blot analysis. Proteins were extracted from cells or tissues and subjected to 4–20% Criterion Precast Midi Protein Gel electrophoresis. Blotted membranes were incubated overnight at 4 °C with primary antibodies, and washed with PBS containing 0.1% Tween-20 before incubation with HRP-conjugated secondary antibody (anti-mouse or anti-rabbit IgG (Promega)) for 1 h followed by chemiluminescent detection (ECL (GE Healthcare)). Antibodies used for western blotting included: rabbit polyclonal anti-AKR1A1 (15054-1-AP, Proteintech Group), rabbit monoclonal anti-PKM1 (D30G6, Cell Signaling), rabbit monoclonal anti-PKM2 (D78A4, Cell Signaling), rabbit polyclonal anti-PKLR (AV41699, Sigma), rabbit

© 2019 Springer Nature Limited. All rights reserved.

polyclonal anti-NOS1 (sc-8309, Santa Cruz Biotechnology), rabbit polyclonal anti-NOS2 (sc-8310, Santa Cruz Biotechnology), rabbit polyclonal anti-NOS3 (sc-654, Santa Cruz Biotechnology), mouse anti-eNOS (pS1177) (612393, BD Transduction Laboratory), mouse monoclonal p97 (10R-P104A, Fitzgerald), rabbit monoclonal GAPDH (Ab181602, abcam), mouse monoclonal Myc (M047-3, MBL), mouse monoclonal Flag–M2 (F3165, Sigma), and mouse monoclonal V5 (R960-25, Invitrogen). Quantification of western blots was carried out with ImageJ (NIH).

Expression of PKM1, PKM2 and PKLR. To measure the expression of PKM1, PKM2 and PKLR after AKI, kidney samples were taken from AKI-injured wild-type mice after 1, 3 or 7 days. To compare the expression of PKM1, PKM2 and PKLR between the kidneys of *Akr1a1*$^{+/+}$ and of *Akr1a1*$^{-/-}$ mice, kidneys were removed after 24 h of AKI. Kidney samples were mechanically homogenized in lysis buffer (50 mM phosphate buffer, pH 7.0, 150 mM NaCl, 0.1 mM EDTA, 1 mM PMSF, and protease inhibitor mixture (Roche)). Extracts were clarified by centrifugation (20,000*g*, 4 °C, 20 min, ×2), and protein concentration was determined by BCA assay. The expression of PKM1, PKM2 and PKLR was examined using western blot.

GSSG/GSH and NADPH/NADP$^+$. The GSSG/GSH ratio was assayed using the GSH/GSSG-Glo Assay kit from Promega. Mouse kidney samples (20 mg) were mechanically homogenized in 100 μl total glutathione lysis reagent for total glutathione measurement or 100 μl oxidized glutathione lysis reagent for GSSG measurement. Extracts were clarified by centrifugation (20,000*g*, 4 °C, 20 min, ×2) and 50 μl supernatant was transferred to the plate reader. Subsequently, 50 μl luciferin generation reagent was added to all wells, and assays were mixed and incubated for 30 min. Then, 100 μl luciferin detection reagent was added to wells followed by mixing. After 15 min of incubation, luminescence was measured using a luminometer. The NADPH/NADP$^+$ assay was carried out using the NADP/NADPH Assay kit from Abcam. Mouse kidney samples (50 mg) were mechanically homogenized in 500 μl of NADP/NADPH extraction buffer. Extracts were clarified by centrifugation (20,000*g*, 4 °C, 10 min) and proteins in the supernatant were removed using Amicon Ultra-0.5 ml Centrifugal Filters. The filtrate was used for assays following the manufacturer's protocol.

Lipid peroxidation. Mouse kidney samples (25 mg) were mechanically homogenized in 250 μl of RIPA buffer (Invitrogen) containing protease inhibitors and 4 μl 2% (w/v) of the lipid antioxidant BHA. After centrifuging the extract at 160*g* for 10 min at 4 °C, the supernatant was used for analysis. For cells, 2×10^7 cells in 1 ml PBS were sonicated on ice for 10 s and the whole homogenate was used in assays. Next, 100 μl of homogenate or 100 μl standard (malondialdehyde) was combined with 10 μl of TCA-TBA-HCl reagent (0.5% (w/v) TBA in 20% (w/v) TCA and 0.33 N HCl) and mixed thoroughly. Then, 1.5 μl 2% (w/v) of the lipid antioxidant BHA was added to prevent lipid peroxidation during the assay. The solution was heated for 15 min in a boiling water bath. After cooling, the flocculent precipitate was removed by centrifugation at 1,000*g* for 10 min. Subseqently, 150 μl sample or standard (in duplicate) was loaded onto the plate reader. The absorbance of the supernatant was measured at 532 nm against a blank that contained reagents minus homogenate. Levels of TBARS (malondialdehyde (MDA) equivalent) were determined with an MDA standard curve.

Metabolomics. Metabolic assays were carried out as described previously[42]. For metabolomic measurements, snap-frozen kidneys were cut to equal weights (20 mg per specimen) and mechanically homogenized into four volumes of ice-cold water. In brief, sugars, sugar phosphates, organic acids, bile acids, nucleotides and other anionic polar metabolites were measured in 30 μl of tissue homogenate using hydrophilic interaction liquid chromatography and multiple reaction monitoring in the negative ion mode on a 5500 QTRAP MS (SCIEX). Amino acids, amines, acylcarnitines, nucleotides, and other cationic polar metabolites were measured in 10 μl of tissue homogenate using hydrophilic interaction liquid chromatography coupled with nontargeted, positive ion mode MS analysis on an Exactive Plus Orbitrap MS (Thermo Scientific). Results were analysed in MetaboAnalyst (http://www.metaboanalyst.ca).

Statistics. Statistics were analysed using Minitab express and GraphPad Prism. Any outliers in data were identified and excluded using GraphPad Prism. Comparisons between continuous characteristics of subject groups were analysed with two-tailed Student's *t*-test using GraphPad Prism. For comparisons among more than two groups, one-way ANOVA with Tukey's post hoc test in Minitab express was used. Survival was analysed by Kaplan–Meier estimation using GraphPad Prism. Overlap of *S*-nitrosylated proteins in three independent experiments (SNO-RAC-coupled quantitative MS) and interactions between the nitrosoproteome and SCoR–AKR1A1 interactome were analysed using the SAS program. No statistical methods were used to predetermine sample size. Mice were randomized to experimental intervention versus control. Each mouse was assigned a code number to enable blinded sham or AKI surgery. Samples were assigned a code number to enable blinded quantitative analyses and measurements including: metabolomics, iTRAQ-coupled LC–MS/MS, SNO/FeNO (using photolysis/chemiluminescence), histology, mitochondrial morphology, serum creatinine and serum BUN. After all data were collected, the results were analysed and decoded. The investigators were not blinded during other experiments and outcome assessments. Bar graphs with the corresponding dot plots were created using GraphPad Prism. Results are presented as mean ± s.d.

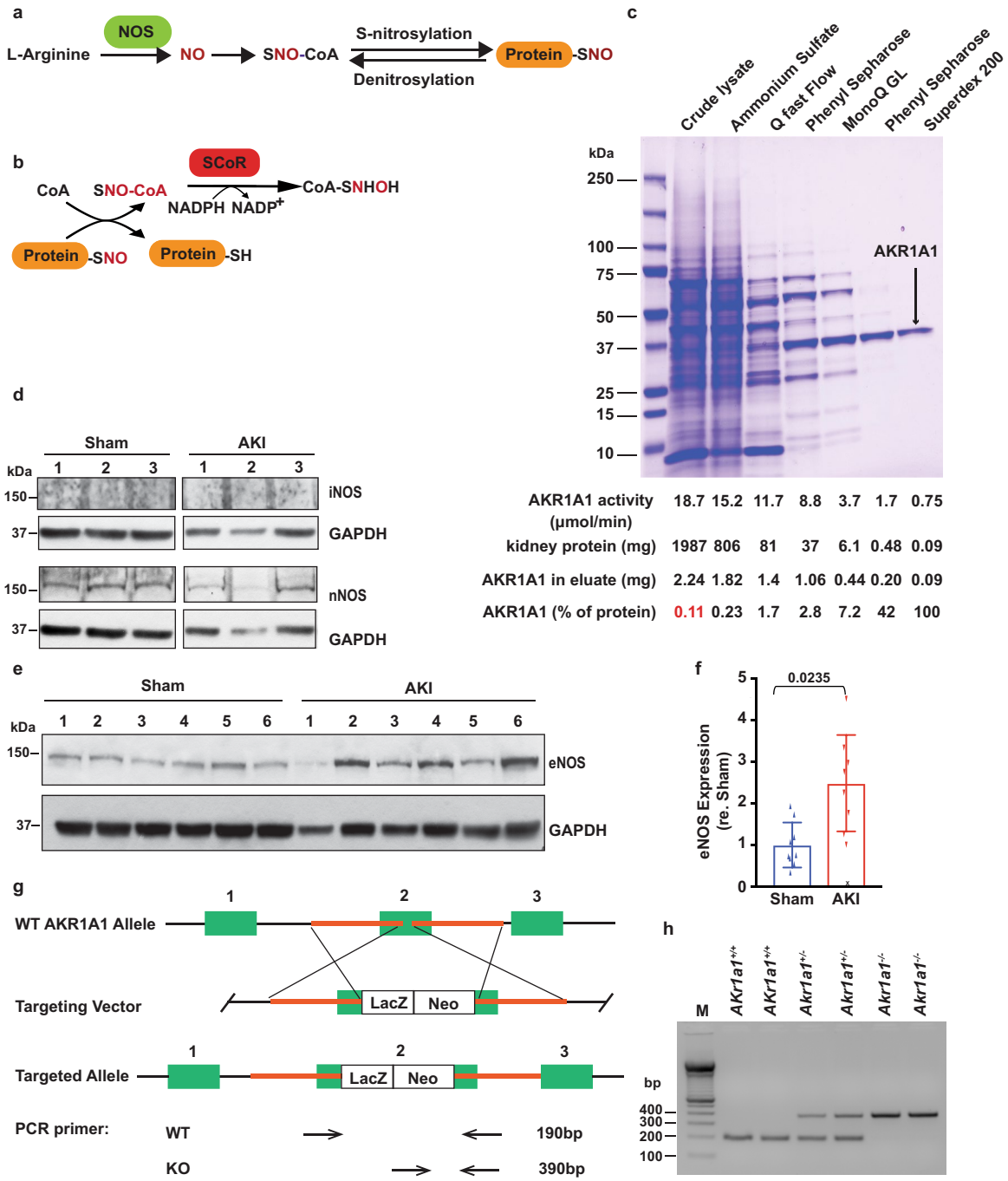
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information. All datasets generated and/or analysed in the current study are available from the corresponding author upon reasonable request. Supplementary Fig. 1 contains scanned complete images of western blots. All experimental data from mice models are provided as Source Data.

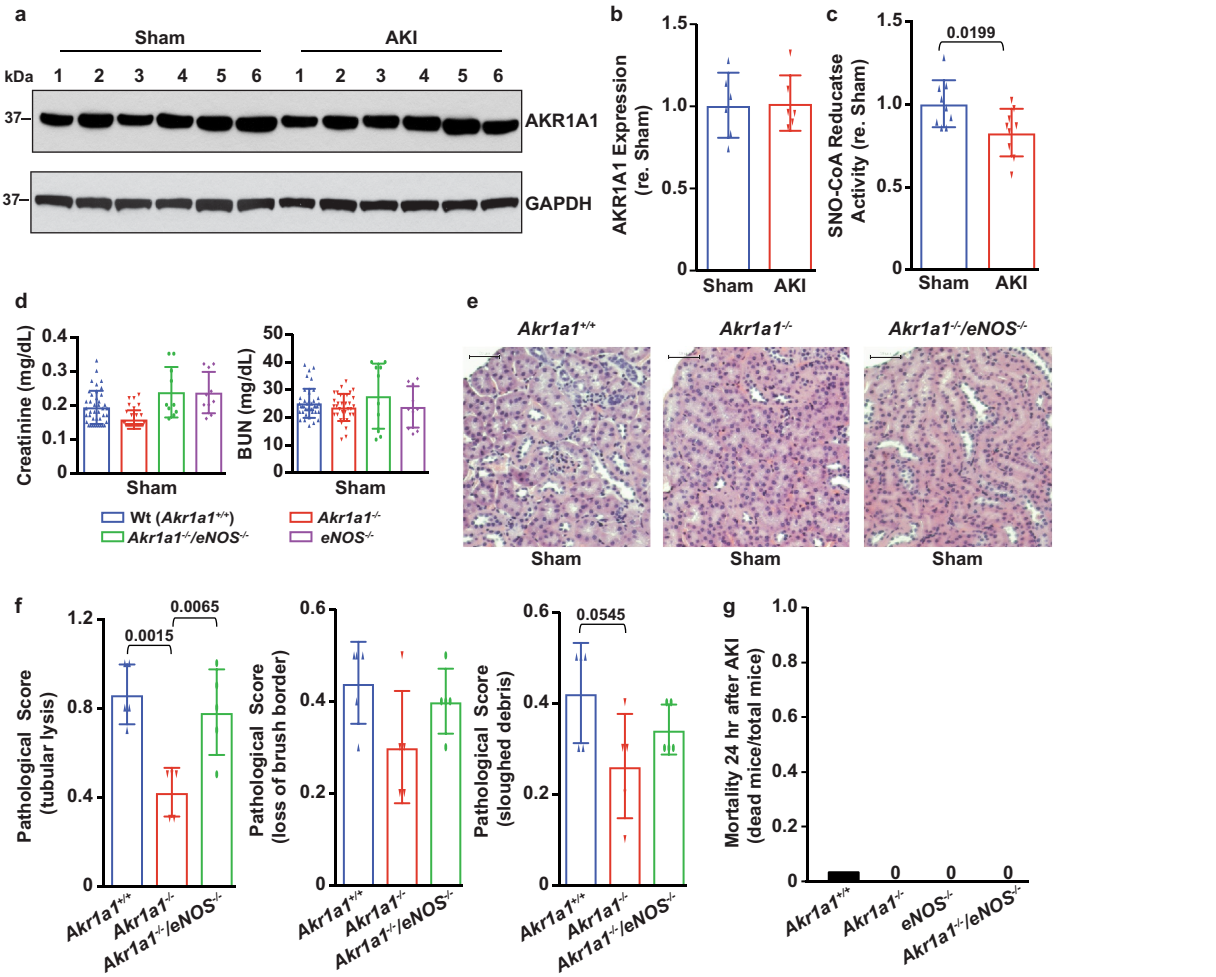
31. Israelsen, W. J. et al. PKM2 isoform-specific deletion reveals a differential requirement for pyruvate kinase in tumor cells. *Cell* **155**, 397–409 (2013).
32. Shao, X., Somlo, S. & Igarashi, P. Epithelial-specific Cre/lox recombination in the developing kidney and genitourinary tract. *J. Am. Soc. Nephrol.* **13**, 1837–1846 (2002).
33. Tran, M. T. et al. PGC1α drives NAD biosynthesis linking oxidative metabolism to renal protection. *Nature* **531**, 528–532 (2016).
34. Forrester, M. T. et al. Proteomic analysis of *S*-nitrosylation and denitrosylation by resin-assisted capture. *Nat. Biotechnol.* **27**, 557–559 (2009).
35. Zhou, H. L., Geng, C., Luo, G. & Lou, H. The p97–UBXD8 complex destabilizes mRNA by promoting release of ubiquitinated HuR from mRNP. *Genes Dev.* **27**, 1046–1058 (2013).
36. Stamler, J. S. et al. Nitric oxide circulates in mammalian plasma primarily as an *S*-nitroso adduct of serum albumin. *Proc. Natl Acad. Sci. USA* **89**, 7674–7677 (1992).
37. Solez, K., Morel-Maroger, L. & Sraer, J. D. The morphology of "acute tubular necrosis" in man: analysis of 57 renal biopsies and a comparison with the glycerol model. *Medicine (Baltimore)* **58**, 362–376 (1979).
38. Conger, J. D., Schultz, M. F., Miller, F. & Robinette, J. B. Responses to hemorrhagic arterial pressure reduction in different ischemic renal failure models. *Kidney Int.* **46**, 318–323 (1994).
39. Fujioka, H., Tandler, B. & Hoppel, C. L. Mitochondrial division in rat cardiomyocytes: an electron microscope study. *Anat. Rec. (Hoboken)* **295**, 1455–1461 (2012).
40. Hanaichi, T. et al. A stable lead by modification of Sato's method. *J. Electron Microsc. (Tokyo)* **35**, 304–306 (1986).
41. Hitosugi, T. et al. Tyrosine phosphorylation inhibits PKM2 to promote the Warburg effect and tumor growth. *Sci. Signal.* **2**, ra73 (2009).
42. Paynter, N. P. et al. Metabolic predictors of incident coronary heart disease in women. *Circulation* **137**, 841–853 (2018).

© 2019 Springer Nature Limited. All rights reserved.



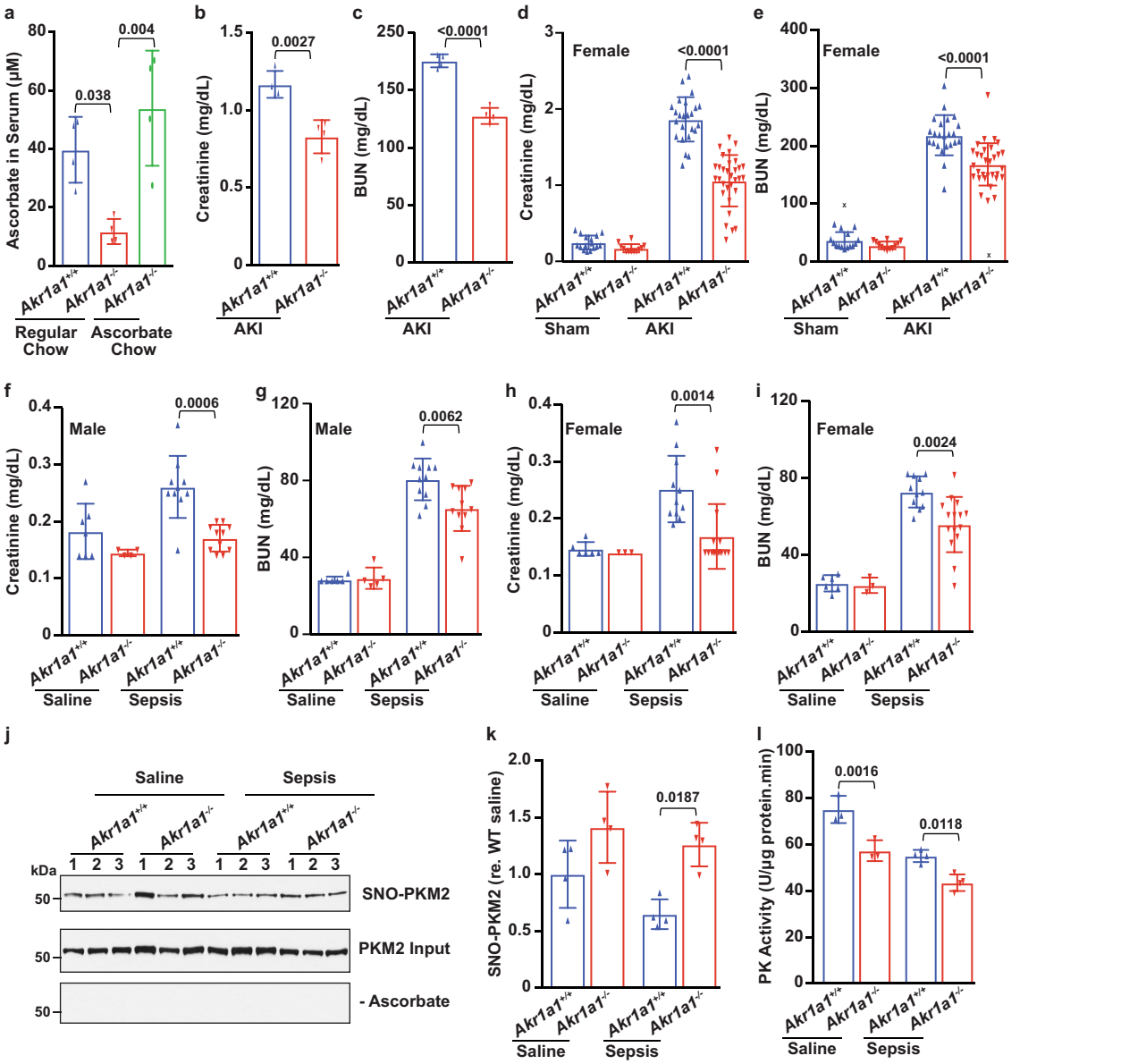
Extended Data Fig. 1 | Identification of enzymes involved in the SNO-CoA–SCoR system. a, **b**, Enzymatic mechanism by which the SNO-CoA–SCoR system regulates protein *S*-nitrosylation. **a**, Equilibrium between SNO-CoA and *S*-nitrosylated proteins. **b**, AKR1A1 (SCoR) mediates protein denitrosylation. **c**, AKR1A1 was purified to homogeneity using the indicated steps. AKR1A1 protein at each stage was calculated on the basis of activity in each eluate pool or original crude lysate. Image is representative of two independently performed experiments with similar results. **d**, **e**, Expression of iNOS, nNOS and eNOS in sham-treated versus injured kidneys of wild-type mice; injury induced by I/R. Images are representative of three independently performed experiments with similar results. For gel source data, see Supplementary Fig. 1. **f**, Expression of eNOS before and after AKI; normalized to GAPDH as in **e** ($n = 8$ mice per group). Results are presented as mean ± s.d. Two-tailed Student's *t*-test was used to detect significance. **g**, Schema illustrating generation of *Akr1a1*$^{-/-}$ mice. **h**, PCR amplification of the *Akr1a1* gene with genomic DNA isolated from the tails of *Akr1a1*$^{+/+}$, heterozygous *Akr1a1*$^{+/-}$ and homozygous *Akr1a1*$^{-/-}$ mice. Image is representative of three independently performed experiments with similar results.

© 2019 Springer Nature Limited. All rights reserved.



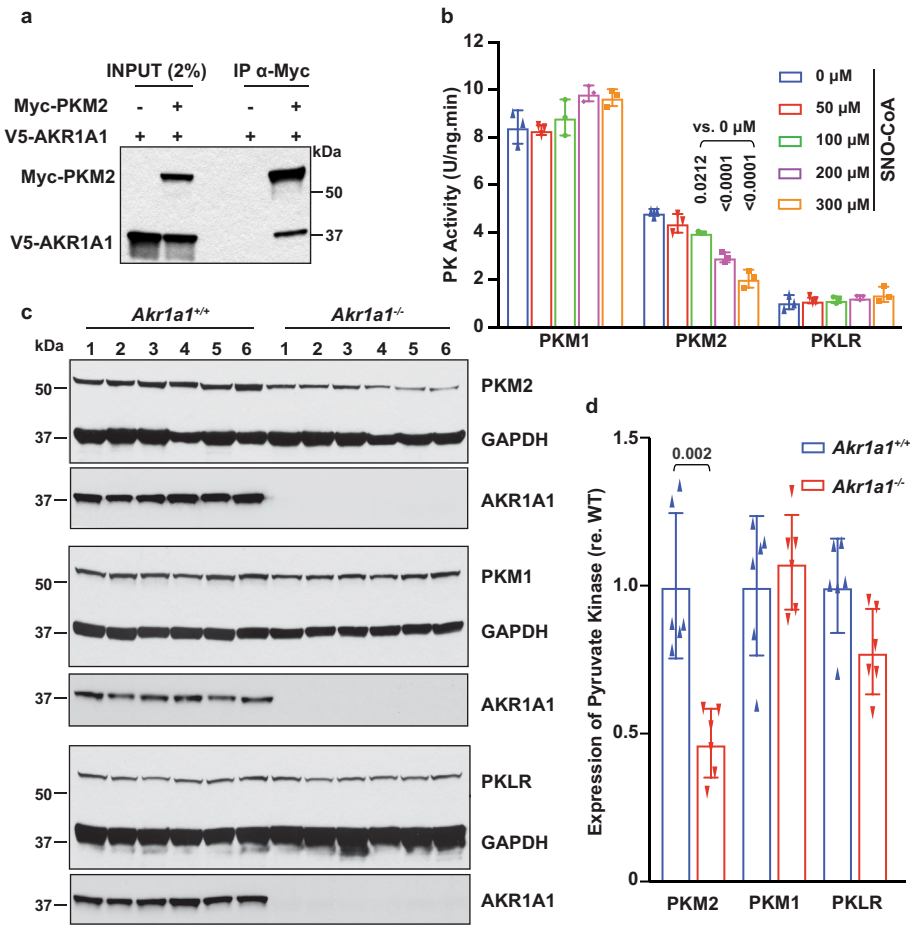
Extended Data Fig. 2 | SCoR activity and role in protection. **a**, **b**, Expression of AKR1A1 after I/R-induced AKI. Expression of AKR1A1 is normalized to GAPDH in **b** ($n = 6$ mice per group). **c**, NADPH-dependent SNO-CoA metabolizing activity measured in kidney extracts from sham-treated wild-type mice or wild-type mice subjected to I/R-induced AKI ($n = 9$ mice per group). **d**, Serum creatinine and BUN in sham-treated kidneys from *Akr1a1*$^{+/+}$, *Akr1a1*$^{-/-}$, *Akr1a1*$^{-/-}$*Enos*$^{-/-}$ and *Enos*$^{-/-}$ mice (*Akr1a1*$^{+/+}$: $n = 41$ mice; *Akr1a1*$^{-/-}$: $n = 35$ mice; *Akr1a1*$^{-/-}$*Enos*$^{-/-}$: $n = 10$ mice; *Enos*$^{-/-}$: $n = 10$ mice). Note lower scales (y axis) compared to Fig. 1e, f. **e**, Haematoxylin and eosin stain of sham-treated kidneys from *Akr1a1*$^{+/+}$, *Akr1a1*$^{-/-}$ and *Akr1a1*$^{-/-}$*Enos*$^{-/-}$ mice. Images are representative of two independently performed experiments with similar results. Scale bars, 50 μm. **f**, Pathological scores of tubular lysis, loss of brush border and sloughed debris ($n = 5$ mice per group). Results are presented as mean ± s.d. One-way ANOVA with Tukey post hoc test was used to detect significance. **g**, Mortality of *Akr1a1*$^{+/+}$, *Akr1a1*$^{-/-}$, *Akr1a1*$^{-/-}$*Enos*$^{-/-}$ and *Enos*$^{-/-}$ mice 24 h after AKI (*Akr1a1*$^{+/+}$: 35 mice; *Akr1a1*$^{-/-}$: 36 mice; *Akr1a1*$^{-/-}$*Enos*$^{-/-}$: 12 mice; *Enos*$^{-/-}$: 8 mice). Results are presented as mean ± s.d. One-way ANOVA with Tukey post hoc test was used to detect significance in **d**, **f**. Two-tailed Student's t-test was used to detect significance in **b**, **c**.

© 2019 Springer Nature Limited. All rights reserved.



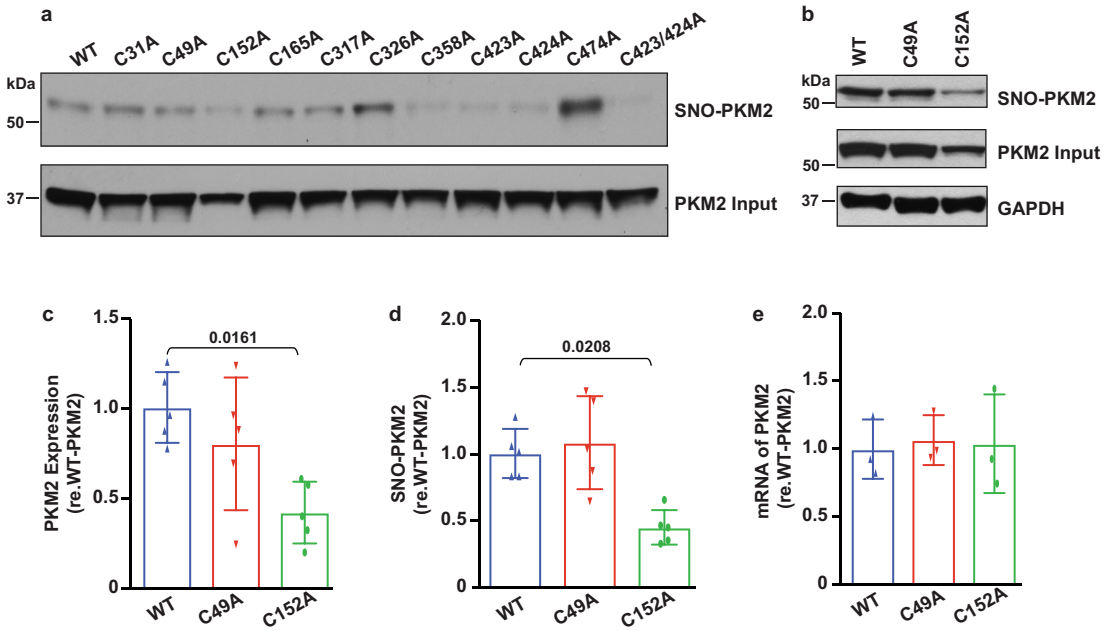
Extended Data Fig. 3 | Additional models of AKI. a, Serum ascorbate in *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice fed with chow containing 1% ascorbic acid for six weeks ($n = 4$ mice per group). **b**, **c**, Serum creatinine and BUN in injured kidneys from *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice fed with chow containing 1% ascorbic acid for six weeks ($n = 4$ mice per group); injury by I/R. **d**, **e**, Serum creatinine and BUN in sham-treated or injured kidneys from female *Akr1a1*$^{+/+}$ and female *Akr1a1*$^{-/-}$ mice (sham-treated *Akr1a1*$^{+/+}$: $n = 11$ mice; sham-treated *Akr1a1*$^{-/-}$: $n = 12$ mice; injured *Akr1a1*$^{+/+}$: $n = 25$ mice; injured *Akr1a1*$^{-/-}$: $n = 31$ mice). Injury by I/R. **f**, **g**, Serum creatinine and BUN in saline-treated or LPS-treated male *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice (saline-treated *Akr1a1*$^{+/+}$: $n = 7$ mice; saline-treated *Akr1a1*$^{-/-}$: $n = 5$ mice; LPS-treated *Akr1a1*$^{+/+}$: $n = 11$ mice; LPS-treated *Akr1a1*$^{-/-}$: $n = 11$ mice). **h**, **i**, Serum creatinine and BUN in saline-treated or LPS-treated female *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice (saline-treated *Akr1a1*$^{+/+}$: $n = 6$ mice; saline-treated *Akr1a1*$^{-/-}$: $n = 3$ mice; LPS-treated *Akr1a1*$^{+/+}$: $n = 12$ mice; LPS-treated *Akr1a1*$^{-/-}$: $n = 15$ mice). **j**, Endogenous *S*-nitrosylation of PKM2 in saline-treated or LPS-treated male *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice. Data are representative of three mice per genotype. Without ascorbate (–Ascorbate) is control for SNO. **k**, Quantification of SNO-PKM2. SNO normalized to PKM2 (input) ($n = 4$ mice per group). **l**, Activity of endogenous pyruvate kinase in saline- or LPS-treated kidneys from *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice ($n = 3$ mice per saline-treated group; $n = 5$ mice per LPS-treated group). Results are presented as mean ± s.d. One-way ANOVA with Tukey post hoc test was used to detect significance **a**, **d**–**i**, **k**, **l**. Two-tailed Student's *t*-test was used to detect significance in **b**, **c**.

© 2019 Springer Nature Limited. All rights reserved.



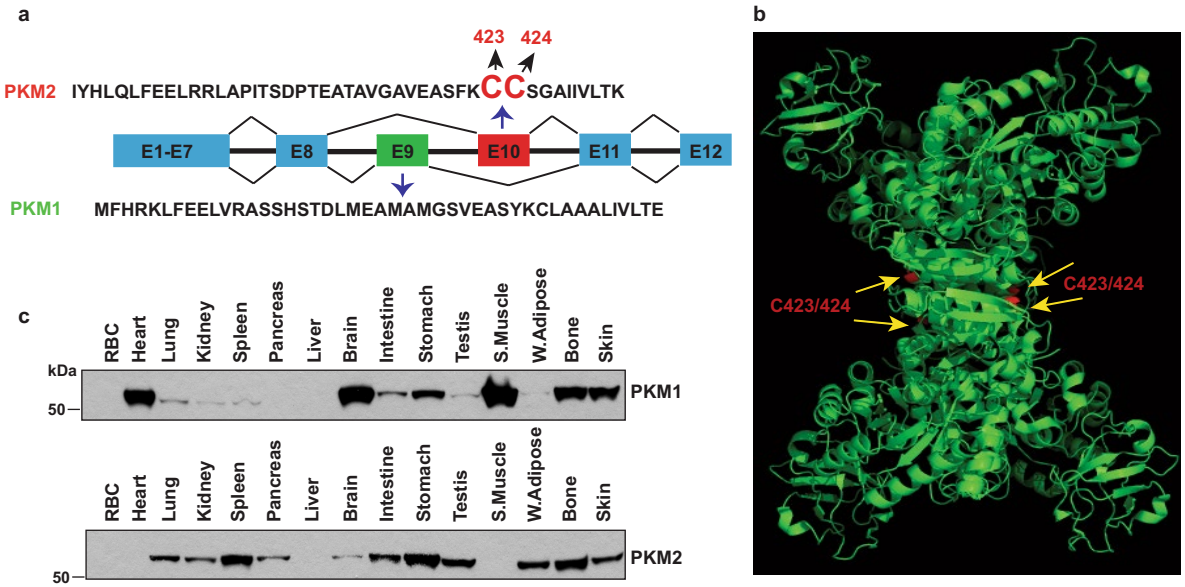
Extended Data Fig. 4 | PK interactions, activity and expression. **a**, Interaction between AKR1A1 and PKM2. Myc–PKM2 and V5–AKR1A1 were co-overexpressed in HEK-293 cells. Immunoprecipitation with anti-Myc rabbit antibody; immunoblotting with V5 antibody. Image is representative of two independently performed experiments with similar results. **b**, Activity of recombinant PKM2, PKM1 and PKLR proteins after SNO-CoA treatment ($n = 3$ independent experiments). Results are presented as mean ± s.d. One-way ANOVA with Tukey post hoc test was used to assess significance. **c**, Expression of PKM2, PKM1 and PKLR in the kidneys of $Akr1a1^{+/+}$ and $Akr1a1^{-/-}$ mice after 24 h of I/R-induced AKI. Image is representative of two independently performed experiments with similar results. **d**, Quantification of expression of PKM2, PKM1 and PKLR in **c** ($n = 6$ mice per group). Result is presented as mean ± s.d. Two-tailed Student's *t*-test was used to detect significance.

© 2019 Springer Nature Limited. All rights reserved.



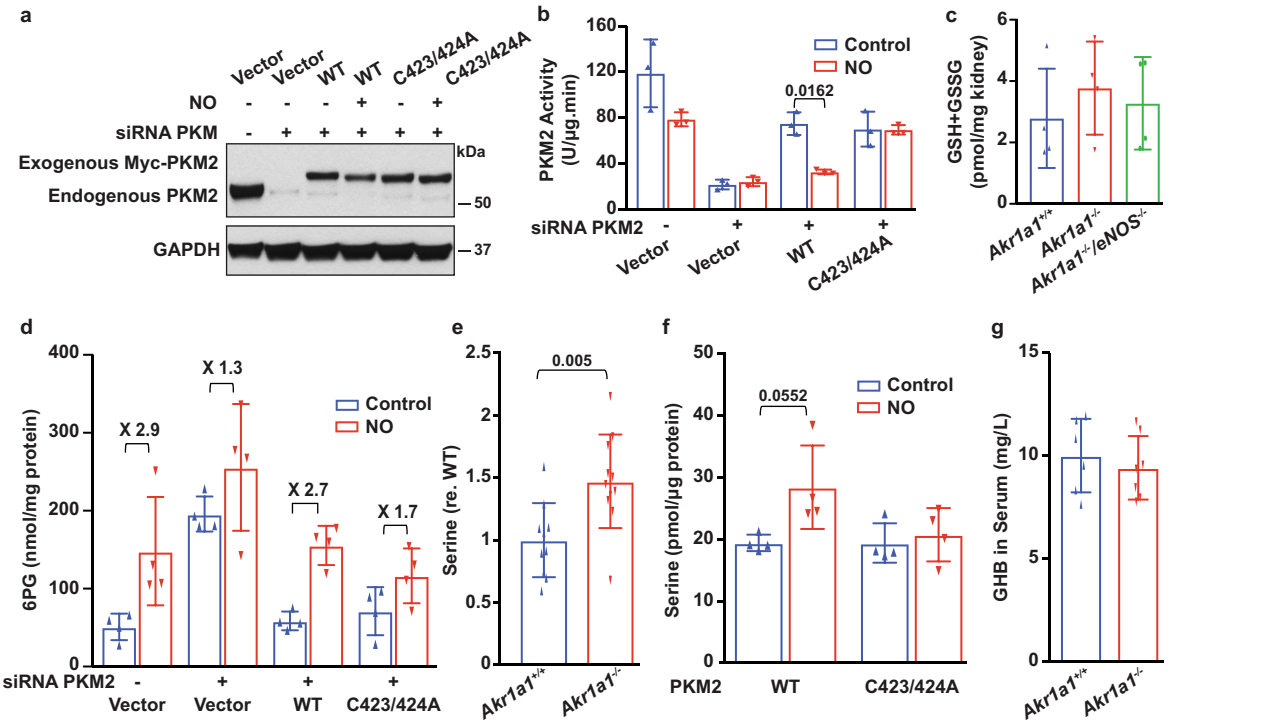
Extended Data Fig. 5 | Characterization of SNO-PKM2. a, Endogenous SNO level of PKM2 Cys-mutants in eNOS-overexpressing HEK-293 cells. Image is representative of two independently performed experiments with similar results. **b**, Mutation of C152 to alanine affects the SNO level of PKM2 in eNOS-overexpressing HEK-293 cells. Image is representative of two independently performed experiments with similar results. **c**, Quantification of expression of Myc–PKM2(WT), Myc–PKM2(C49A) and Myc–PKM2(C152A) in eNOS-overexpressing HEK-293 cells. Normalized to the expression of GAPDH ($n = 5$ independent experiments). **d**, Quantification of SNO-PKM2 in eNOS-overexpressing HEK-293 cells. SNO is normalized to PKM2 (input) ($n = 5$ independent experiments). **e**, Relative mRNA levels of Myc–PKM2(WT), Myc–PKM2(C49A) and Myc–PKM2(C152A) in eNOS-overexpressing HEK-293 cells ($n = 3$ independent experiments). Results presented as mean ± s.d. One-way ANOVA with Tukey post hoc test was used to detect significance.

© 2019 Springer Nature Limited. All rights reserved.



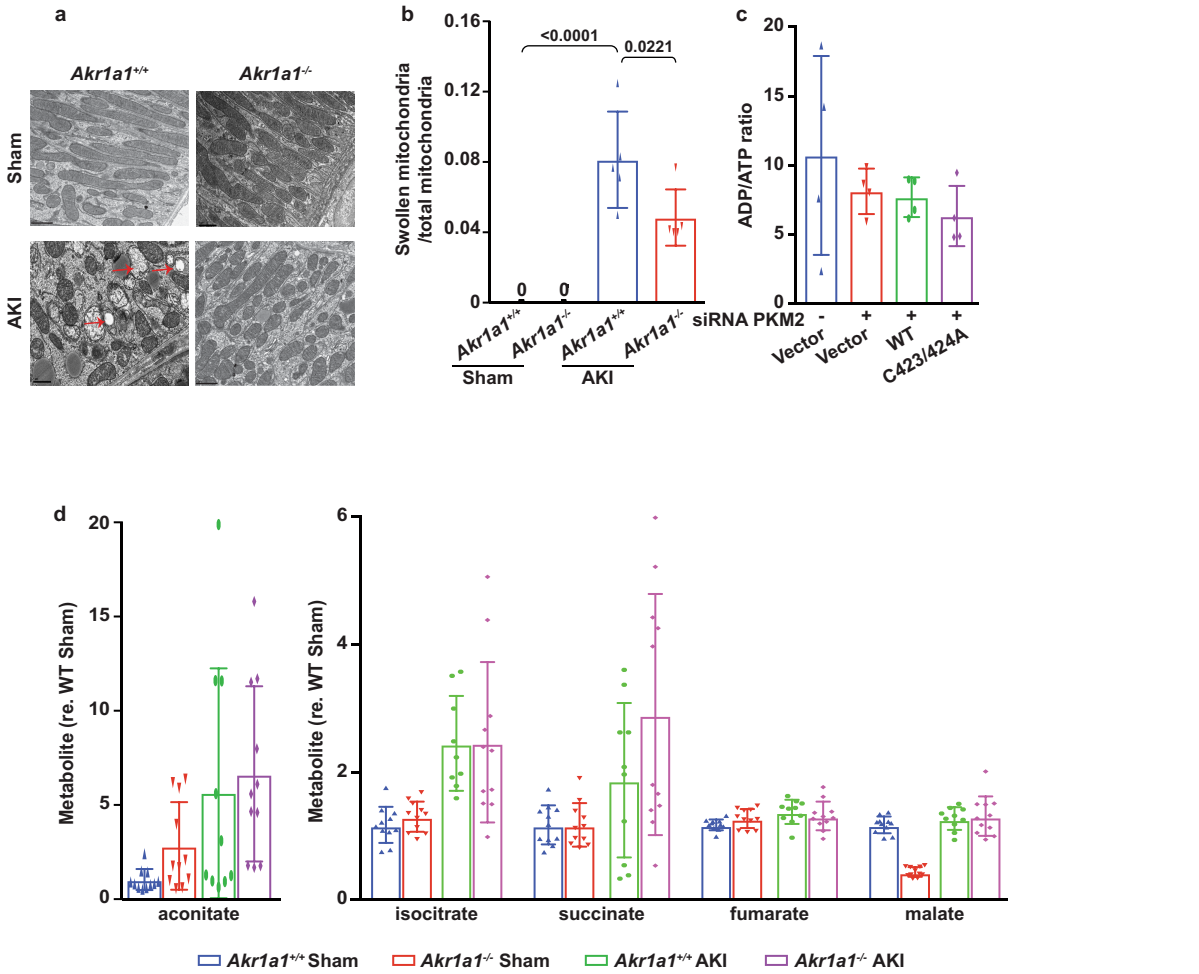
Extended Data Fig. 6 | *Pkm* gene, structure and expression. **a**, Alternative splicing of *Pkm* gene. C423 and C424 are encoded by PKM2-specific exon 10. **b**, Ribbon structure of tetrameric PKM2 analysed by MacPyMOL. Four pairs of C423 and C424 in tetrameric PKM2 are highlighted in red. **c**, Expression of PKM1 and PKM2 in fifteen different tissues from uninjured mice. Image is representative of two independently performed experiments with similar results.

© 2019 Springer Nature Limited. All rights reserved.



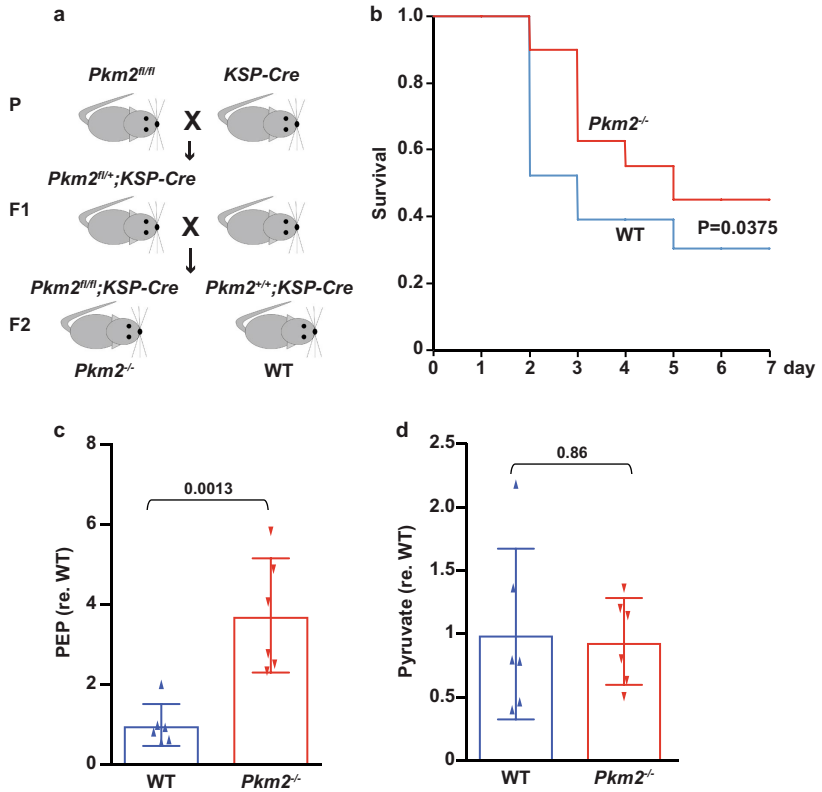
Extended Data Fig. 7 | Role of PKM2 in AKR1A1-mediated protection. a, Expression of endogenous and overexpressed PKM2 in HEK-293 cells. Image is representative of two independently performed experiments with similar results. **b**, Activity of Myc–PKM2(WT) and Myc–PKM2(C423/424A) after NO (DETANO; 500 µM) treatment in HEK-293 cells ($n = 3$ independent experiments). **c**, The total amount of GSH and GSSG in injured kidneys from *Akr1a1*$^{+/+}$, *Akr1a1*$^{-/-}$ and *Akr1a1*$^{-/-}$*Enos*$^{-/-}$ mice ($n = 4$ mice per group); injury induced by I/R. **d**, The amount of 6PG, a key PPP intermediate, in HEK-293 cells expressing Myc–PKM2(WT) and Myc–PKM2(C423/424A) after NO (DETANO; 500 µM) treatment ($n = 4$ independent experiments). **e**, Amount of serine in injured kidneys from *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice ($n = 10$ mice per I/R-injured *Akr1a1*$^{+/+}$ group; $n = 11$ mice per I/R-injured *Akr1a1*$^{-/-}$ group). **f**, Amount of serine in HEK-293 cells expressing Myc–PKM2(WT) and Myc–PKM2(C423/424A) after NO treatment (DETANO; 500 µM) ($n = 4$ independent experiments). **g**, Amount of GHB in serum of *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice ($n = 7$ mice per group). No injury. Results are presented as mean ± s.d. One-way ANOVA with Tukey post hoc test was used to detect significance in **b** and **f**. Two-tailed Student's *t*-test was used in **e**.

© 2019 Springer Nature Limited. All rights reserved.



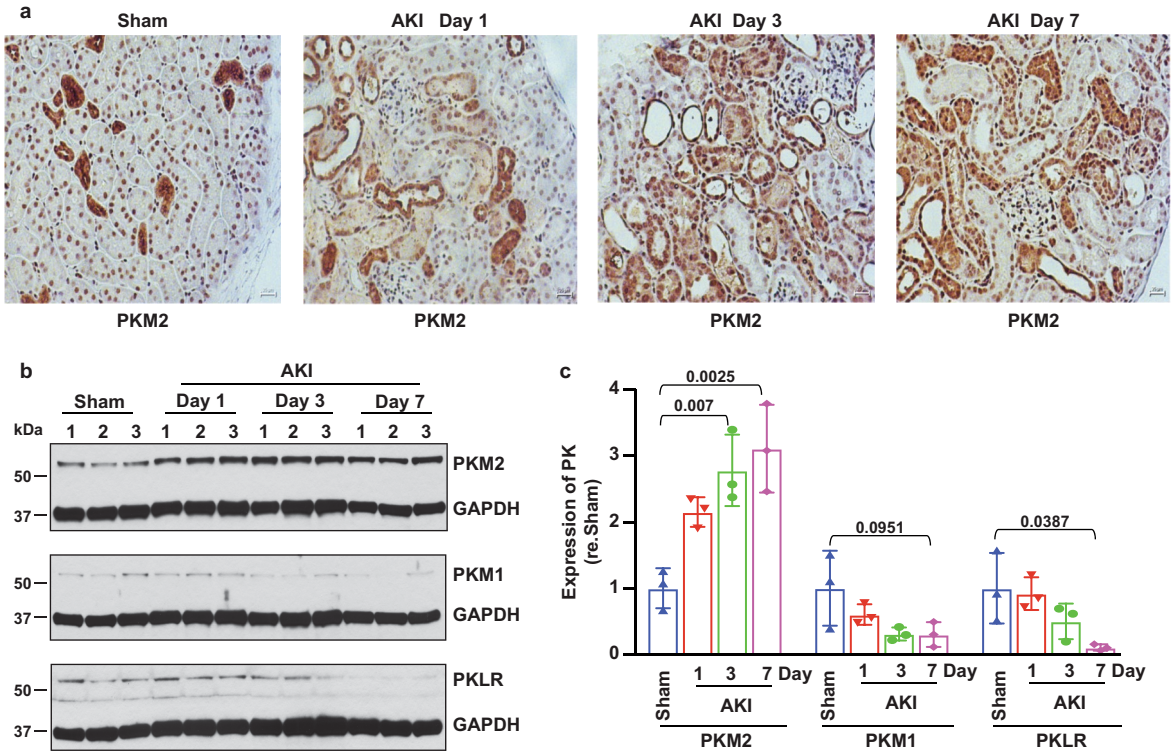
Extended Data Fig. 8 | Mechanism of kidney injury. a, Mitochondrial morphology in tubule cells after sham operation or injury induced by I/R from *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice as assessed by electron microscopy. Red arrows, mitochondrial swelling. Scale bars, 1 μm. **b**, Quantification of swollen mitochondria versus total mitochondria in sham-treated or I/R-injured kidneys from *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice. Results are presented as mean ± s.d. One-way ANOVA with Tukey post hoc was used to detect significance. **c**, The ratio of ADP to ATP in HEK-293 cells expressing Myc–PKM2(WT) or Myc–PKM2(C423/424A) after NO treatment (DETANO; 500 μM) ($n = 4$ independent experiments). **d**, Amounts of TCA cycle intermediates (aconitate, isocitrate, succinate, fumarate and malate) in sham-treated or injured kidneys from *Akr1a1*$^{+/+}$ and *Akr1a1*$^{-/-}$ mice ($n = 10$ mice per sham-treated *Akr1a1*$^{+/+}$ or I/R-injured *Akr1a1*$^{+/+}$ group; $n = 11$ mice per sham-treated *Akr1a1*$^{-/-}$ or I/R-injured *Akr1a1*$^{-/-}$group). Results are presented as mean ± s.d. There were no significant differences in **c**, **d** using one-way ANOVA with Tukey's post hoc test.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Characterization of *Pkm2*$^{-/-}$ mice. **a**, Schema illustrating generation of renal epithelial cell-specific *Pkm2*$^{-/-}$ mice. **b**, Survival curve following I/R-induced AKI (23 wild-type mice; 20 *Pkm2*$^{-/-}$ mice). Survival was analysed using Kaplan–Meier estimation. Gehan–Breslow–Wilcoxon test was used to detect significance. **c**, PEP in injured kidneys from wild-type and *Pkm2*$^{-/-}$ mice ($n = 6$ mice per group); injury induced by I/R. **d**, Pyruvate in injured kidneys from wild-type and *Pkm2*$^{-/-}$ mice ($n = 6$ mice per group). Results in **c**, **d** are presented as mean ± s.d. Two-tailed Student's *t*-test was used to detect significance.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | Expression of PKM1, PKM2 and PKLR after AKI. a, Immunostaining showing expression of PKM2 in sham or injured kidneys of wild-type mice on the indicated days after surgery; injury induced by I/R. Images are representative of two independently performed experiments with similar results. **b**, Western blot showing expression of PKM2, PKM1 and PKLR in sham or injured kidneys of wild-type mice; injury induced by I/R. Images are representative of two independently performed experiments with similar results. **c**, Quantification of expression of PKM2, PKM1 and PKLR in **b** ($n = 3$ mice). Results are presented as mean ± s.d. One-way ANOVA with Tukey's post hoc test was used to detect significance.

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0806-7

Metabolic heterogeneity underlies reciprocal fates of T_H17 cell stemness and plasticity

Peer W. F. Karmaus[1], Xiang Chen[2], Seon Ah Lim[1], Andrés A. Herrada[1], Thanh-Long M. Nguyen[1], Beisi Xu[2], Yogesh Dhungana[1], Sherri Rankin[1], Wenan Chen[2], Celeste Rosencrance[2], Kai Yang[1], Yiping Fan[2], Yong Cheng[3], John Easton[2], Geoffrey Neale[4], Peter Vogel[5] & Hongbo Chi[1]*

A defining feature of adaptive immunity is the development of long-lived memory T cells to curtail infection. Recent studies have identified a unique stem-like T-cell subset amongst exhausted CD8-positive T cells in chronic infection[1–3], but it remains unclear whether CD4-positive T-cell subsets with similar features exist in chronic inflammatory conditions. Amongst helper T cells, T_H17 cells have prominent roles in autoimmunity and tissue inflammation and are characterized by inherent plasticity[4–7], although how such plasticity is regulated is poorly understood. Here we demonstrate that T_H17 cells in a mouse model of autoimmune disease are functionally and metabolically heterogeneous; they contain a subset with stemness-associated features but lower anabolic metabolism, and a reciprocal subset with higher metabolic activity that supports transdifferentiation into T_H1-like cells. These two T_H17-cell subsets are defined by selective expression of the transcription factors TCF-1 and T-bet, and by discrete levels of CD27 expression. We also identify signalling via the kinase complex mTORC1 as a central regulator of T_H17-cell fate decisions by coordinating metabolic and transcriptional programmes. T_H17 cells with disrupted mTORC1 signalling or anabolic metabolism fail to induce autoimmune neuroinflammation or to develop into T_H1-like cells, but instead upregulate TCF-1 expression and acquire stemness-associated features. Single-cell RNA sequencing and experimental validation reveal heterogeneity in fate-mapped T_H17 cells, and a developmental arrest in the T_H1 transdifferentiation trajectory upon loss of mTORC1 activity or metabolic perturbation. Our results establish that the dichotomy of stemness and effector function underlies the heterogeneous T_H17 responses and autoimmune pathogenesis, and point to previously unappreciated metabolic control of plasticity in helper T cells.

We hypothesized that T_H17 cells in autoimmune microenvironments are heterogeneous and consist of subpopulations with differing extents of lineage stability and plasticity. In the transcriptome of T_H17 cells from mice with experimental autoimmune encephalomyelitis (EAE)[8], *Cd27* (ref. [9]) was more highly expressed in cells from draining lymph nodes than those from the central nervous system (CNS; Extended Data Fig. 1a). We used a fate-mapping system—crossing Il17aCre mice with animals expressing the $R26R^{eYFP}$ reporter—in order to permanently mark those cells that had expressed interleukin (IL)-17 with yellow fluorescent protein (YFP)[6]. Following immunization with myelin oligodendrocyte glycoprotein peptide (MOG), CD27 showed a unique bimodal expression, with distinct $CD27^+$ and $CD27^-$ fractions observed amongst YFP^+ T_H17 cells (Extended Data Fig. 1b). As EAE progressed, the $CD27^+$ subset initially shrank in proportion and then stabilized (Extended Data Fig. 1c), and at the peak of disease it was detected mainly in lymphoid tissues (draining lymph nodes and spleen) but not the spinal cord (Fig. 1a). Amongst YFP^+ T_H17 cells isolated from mice at day nine post-immunization (used throughout this study, unless otherwise noted), the $CD27^+$ fraction expressed decreased levels of IL-17, interferon (IFN)-γ (Fig. 1b) and the proliferative marker Ki-67 (Extended Data Fig. 1d). After ex vivo MOG stimulation, $CD27^+$ cells proliferated and converted into $CD27^-$ cells, whereas $CD27^-$ cells remained $CD27^-$ (Fig. 1c). When transferred into naive hosts, again a fraction of $CD27^+$ cells developed into $CD27^-$ cells, while $CD27^-$ YFP^+ cells remained $CD27^-$ (Extended Data Fig. 1e). Moreover, $CD27^+$ cells expressed high levels of T-cell factor (TCF)-1 and Bcl-2, proteins that mediate $CD8^+$ T cell memory[10,11] (Fig. 1d and Extended Data Fig. 1f). Persistence of $CD27^+$ cells was also enhanced when they were transferred into recipients lacking the Rag1 protein (which is required for lymphocyte development) (Fig. 1e). Therefore, $CD27^+$ cells are characterized by in vivo quiescence and persistence, and by the ability to develop into $CD27^-$ cells.

To explore the underlying cellular and functional processes, we performed transcriptome analysis of $CD27^+$ and $CD27^-$ cells (Extended Data Fig. 2a). Gene-set-enrichment analysis (GSEA) using effector and memory-precursor $CD8^+$ T cell signatures[12] and $CD8^+$ memory and T_{FH} (follicular helper) overlapping signatures[13] suggested that the $CD27^+$ subset was highly enriched for memory-associated signatures, while the $CD27^-$ subset was enriched for genes that are related to effector cells or downregulated in memory cells (Fig. 1f and Extended Data Fig. 2b). The transcriptomes of $CD27^+$ and $CD27^-$ T_H17 cells also showed considerable similarity with those of memory-like $CXCR5^+$ and $CXCR5^-$ subsets, respectively, of exhausted $CD8^+$ T cells[1–3] (Fig. 1f and Extended Data Fig. 2b). Using 'hallmark' gene sets in GSEA, we found that the $CD27^-$ subset was enriched for hallmarks of apoptosis (Extended Data Fig. 2c), mTORC1 signalling, Myc targets, and metabolic pathways including cholesterol homeostasis and glycolysis (Extended Data Fig. 2d). Consistent with these results, mTORC1 activity and Myc expression were higher in $CD27^-$ than in $CD27^+$ T_H17 cells (Fig. 1g, h). Moreover, blocking glycolysis with 2-deoxyglucose (2-DG) increased CD27 expression in YFP^+ cells (Fig. 1i), indicating stabilization of the $CD27^+$ T_H17 phenotype. Overall, T_H17 cells are heterogeneous and consist of two subpopulations, with the $CD27^+$ subset showing memory-like features and metabolic quiescence.

Emerging studies have highlighted the importance of metabolic reprogramming to T-cell activation and differentiation[14], but metabolic control of cellular plasticity is unclear. We therefore determined the effects on T_H17 heterogeneity and plasticity of blocking mTORC1 activity by using the Il17aCre and $R26R^{eYFP}$ system[6] to delete the signature mTORC1 component Raptor (encoded by *Rptor*). Naive T cells from wild-type and $Rptor^{Il17aCre}$ mice showed comparable proliferation and T_H17 differentiation (Extended Data Fig. 3a). However, $Rptor^{Il17aCre}$ mice were protected from MOG-induced EAE (Fig. 2a), and exhibited lack of CNS inflammation and T-cell infiltration (Fig. 2b and Extended Data Fig. 3b). Raptor-deficient YFP^+ cells from the CNS upregulated IL-17 but were defective in IFN-γ expression (Fig. 2c). Thus, mTORC1 function, beyond initial T_H17 differentiation, is crucial for EAE and robust IFN-γ expression.

[1]Department of Immunology, St Jude Children's Research Hospital, Memphis, TN, USA. [2]Department of Computational Biology, St Jude Children's Research Hospital, Memphis, TN, USA. [3]Department of Hematology, St Jude Children's Research Hospital, Memphis, TN, USA. [4]Hartwell Center for Bioinformatics and Biotechnology, St Jude Children's Research Hospital, Memphis, TN, USA. [5]Department of Pathology, St Jude Children's Research Hospital, Memphis, TN, USA. *e-mail: hongbo.chi@stjude.org

© 2019 Springer Nature Limited. All rights reserved.

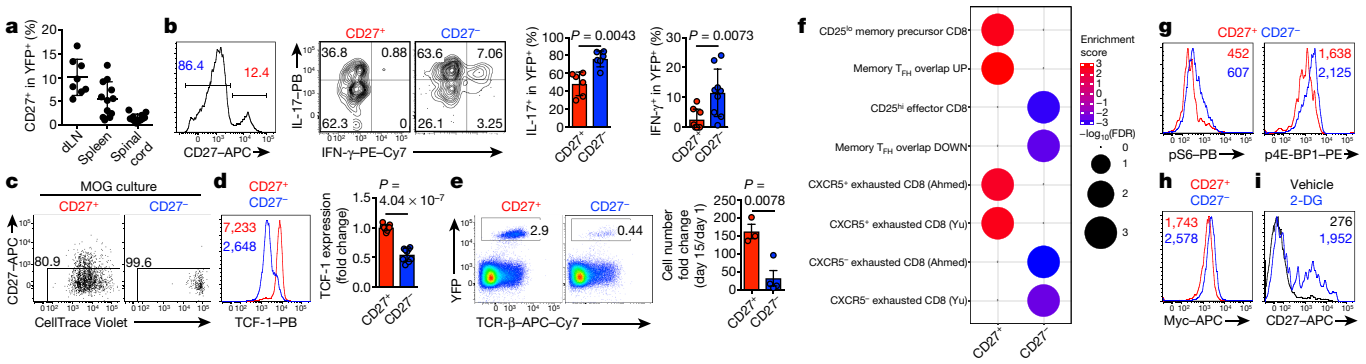


Fig. 1 | CD27$^+$ T$_H$17 cells have memory-like features and low metabolic activity. **a**, Summary of CD27 expression on CD4$^+$ TCR-β$^+$ YFP$^+$ cells at day 16 after MOG immunization in draining lymph nodes (dLN), spleen and spinal cord of Il17aCre (R26R^{eYFP}) mice ($n = 8$, dLN; $n = 12$, spleen and spinal cord). **b**–**h**, Analysis of CD27$^+$ and CD27$^-$ YFP$^+$ populations (**b**, left) from Il17aCre (R26R^{eYFP}) mice at day nine after MOG immunization. **b**, Centre and right panels, IL-17 and IFN-γ expression ($n = 6$, CD27$^+$/CD27$^-$ IL-17; $n = 8$, CD27$^+$ IFN-γ; $n = 9$, CD27$^-$ IFN-γ). **c**, In vitro culture with MOG for analyses of proliferation (CellTrace Violet) and CD27 expression. **d**, TCF-1 expression in CD27$^+$ and CD27$^-$ cells (left) and fold change (right; expression in CD27$^+$ population was set to 1) ($n = 9$). **e**, CD27$^+$ or CD27$^-$ YFP$^+$ cells were transferred into *Rag1*$^{-/-}$ mice, and analysed at day 15 for donor-cell percentages (left) and numbers (right, normalized against cell numbers at day 1) ($n = 3$, CD27$^+$; $n = 4$, CD27$^-$). **f**, GSEA using gene sets related to T-cell memory from acute (top four panels) and chronic (bottom four panels) infection. UP and DOWN indicate upregulated and downregulated genes, respectively, that are shared between memory CD8$^+$ and T$_{FH}$ cells[13], and Ahmed and Yu refer to refs [1,3]. **g**, **h**, Flow cytometry of phosphorylated S6 and 4E-BP1 proteins (**g**) and Myc protein (**h**). **i**, CD27 expression on CD4$^+$ TCR-β$^+$ YFP$^+$ cells stimulated with MOG and vehicle or 2-deoxyglucose (2-DG). Numbers within histograms represent mean fluorescence intensities. Data are means ± s.e.m.; Mann–Whitney U-test (two-sided) in **b**; Student's *t*-test (two-sided) in **d**, **e**. Data are representative of three (**a**, **e**, **g**, **i**), four (**b**–**d**), one (**f**) or two (**h**) independent experiments. FDR is the false discovery rate; PE–Cy7, PB, PE and APC, and APC–Cy7 are different fluorescent labels.

Because IL-17 can be produced by cells other than T$_H$17 cells, we constructed mixed bone-marrow chimaeras to restrict Raptor deficiency to αβ T$_H$17 cells (Extended Data Fig. 3c). *Rptor*^{Il17aCre}-derived chimaeras were resistant to EAE (Extended Data Fig. 3c), indicating the role of mTORC1 in T$_H$17 cells selectively. Additionally, to exclude the effects of absent inflammation in *Rptor*^{Il17aCre} mice, we generated mixed chimaeras using CD45.2$^+$ *Rptor*^{Il17aCre} (or wild-type control) and CD45.1$^+$ wild-type bone-marrow-derived cells, which would mediate CNS inflammation in EAE (Extended Data Fig. 3d). Following MOG immunization, Raptor-deficient cells showed reduced cellularity in CNS, with a preferential loss of IL-17$^-$ IFN-γ$^+$ cells (Extended Data Fig. 3e, f). Raptor-deficient cells also had reduced expression of Ki-67 and T-bet but normal survival and expression of the CCR6 and CXCR3 chemokine receptors (Extended Data Fig. 3g–k). These results identify an intrinsic requirement for mTORC1 in mediating IFN-γ production and in sustaining T$_H$17 responses at the site of inflammation.

In draining lymph nodes from MOG-immunized *Rptor*^{Il17aCre} (R26R^{eYFP}) mice, we found a modestly lower percentage of YFP$^+$ cells, which showed efficient *Rptor* deletion and diminished mTORC1 activity (Extended Data Fig. 4a, b). Raptor-deficient YFP$^+$ cells exhibited normal survival, chemokine receptors and IL-17 expression, but produced less IFN-γ (Extended Data Fig. 4c–e). Moreover, Raptor-deficient cells had reduced expression of T-bet, *Tbx21* and *Il12rb2*, normal expression of ROR-γt and Foxp3, and elevated expression of *Rorc* and *Il23r* (Fig. 2d and Extended Data Fig. 4f, g). Thus, loss of Raptor in T$_H$17 cells impairs expression of T$_H$1-associated factors. In response to ex vivo MOG stimulation, Raptor-deficient T$_H$17 cells produced much less IFN-γ and modestly more IL-17 (Fig. 2e), with largely unaffected proliferation (Extended Data Fig. 4h). Addition of IL-12 converted many IL-17-producing cells into IL-17$^-$ IFN-γ$^+$ cells, but Raptor-deficient cells were less capable of such conversion or transdifferentiation (Fig. 2f). Conversely, IL-23 induced a predominant IL-17$^+$ IFN-γ$^+$ population in a Raptor-dependent manner (Extended Data Fig. 4i). Whereas both cytokines induced IFN-γ expression, IL-23 but not IL-12 maintained IL-17 expression, suggesting a more complete T$_H$1 transdifferentiation induced by IL-12. Furthermore, Raptor-deficient cells were impaired in IL-12-induced phosphorylation of the

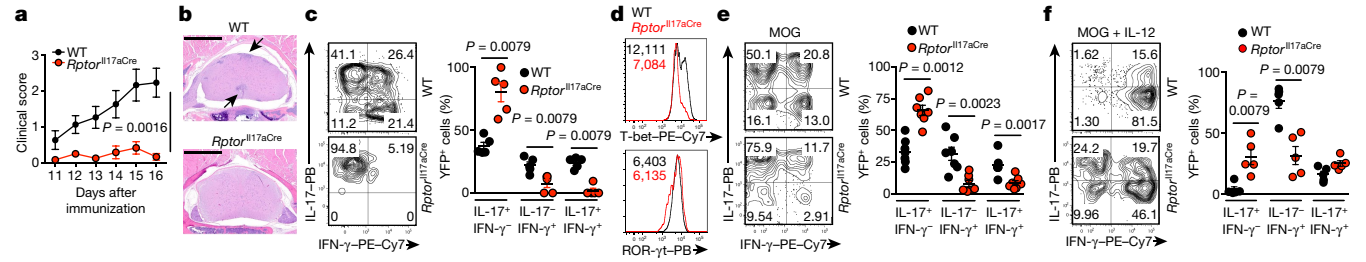


Fig. 2 | Deletion of Raptor in T$_H$17 cells diminishes autoimmune pathogenesis and transdifferentiation into T$_H$1-like cells. Wild-type (WT) and *Rptor*^{Il17aCre} (R26R^{eYFP}) mice were immunized with MOG. **a**, Clinical scores ($n = 15$, WT; $n = 12$, *Rptor*^{Il17aCre}). Normal mouse; 0, no overt signs of disease; 1, limp tail; 2, limp tail plus hindlimb weakness; 3, total hindlimb paralysis; 4, hindlimb paralysis plus 75% of body paralysis (forelimb paralysis/weakness); 5, moribund. **b**, Histopathology of spinal-cord sections at day 16 post-immunization. Scale bars represent 1 mm; arrows indicate lesions. **c**, Flow cytometry analysis (left) and summary (right) of cytokine expression within YFP$^+$ cells from spinal cord at day 16 post-immunization ($n = 5$ per genotype). **d**, T-bet and ROR-γt expression in YFP$^+$ cells from draining lymph nodes at day nine post-immunization. **e**, **f**, Cytokine production by YFP$^+$ cells from draining lymph nodes after four days of stimulation with MOG (**e**; $n = 7$ per genotype) or MOG plus IL-12 (**f**; $n = 5$ per genotype). Numbers within histograms represent mean fluorescence intensities. Data are means ± s.e.m.; two-way analysis of variance (ANOVA) in **a**; Mann–Whitney U-test (two-sided) in **c**, **e**, **f**. Data are pooled from three experiments (**a**), or representative of three (**b**–**d**), seven (**e**) or five (**f**) independent experiments.

© 2019 Springer Nature Limited. All rights reserved.

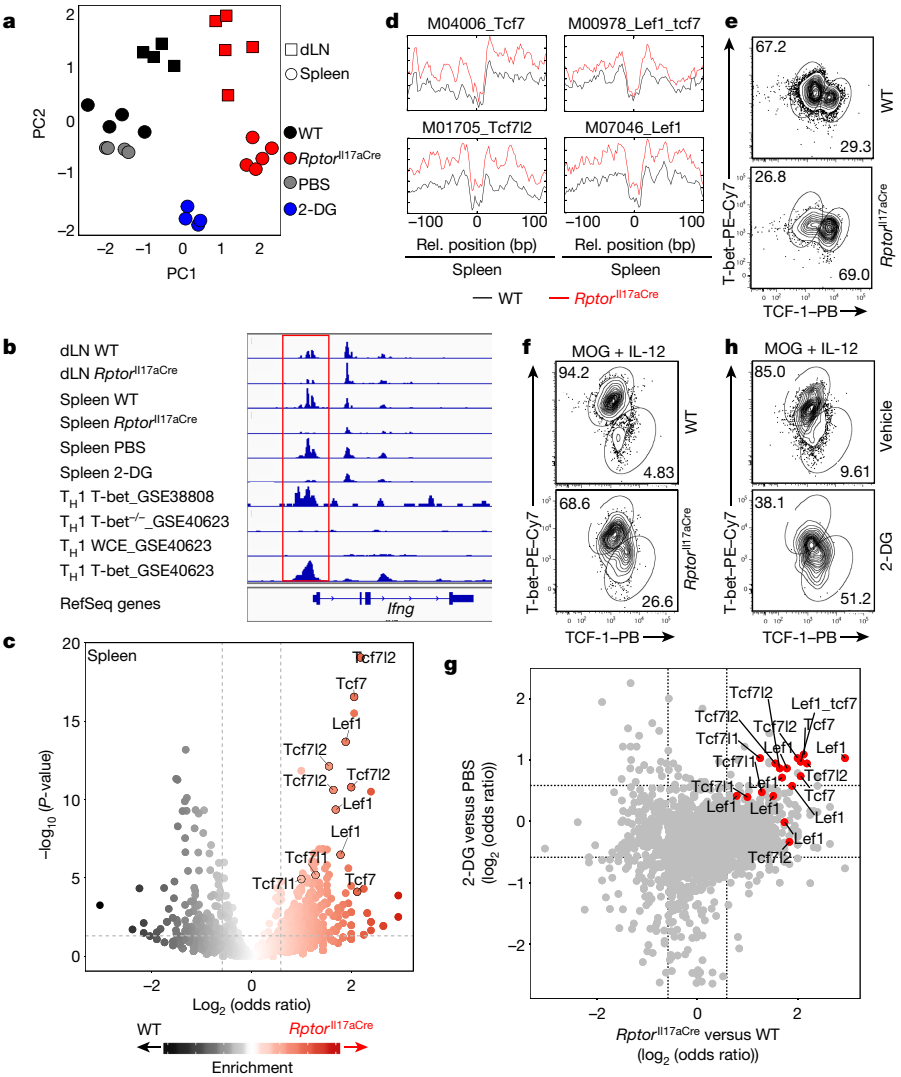


Fig. 3 | TCF-1 expression is regulated by mTORC1 and metabolic activity in T_H17 cells. a–d, ATAC-seq analysis of YFP$^+$ cells isolated from the draining lymph nodes (dLN) and spleen of WT and *Rptor*^{Il17aCre} (R26R^{eYFP}) mice, or the spleen of R26R^{eYFP} mice treated with phosphate-buffered saline (PBS) vehicle or 2-DG at day nine post MOG-immunization. **a**, Principal component analysis (PCA) plot of nucleosome-free fragments. **b**, Accessibility of the *Ifng* locus in different animals, aligned with T-bet-binding sites (red boxes show promoter regions). The bottom four rows show T-bet-binding signals in the indicated public database (WCE, whole-cell extract, control sample without immunoprecipitation). **c**, Motif-enrichment analysis of ATAC-seq from the spleen. **d**, Binding profiles of selected transcription factors from the TCF–LEF family. 'M' numbers refer to TRANSFAC accession numbers for the respective motif sequences. **e**, **f**, Flow cytometry of T-bet and TCF-1 expression in freshly isolated YFP$^+$ cells from dLN of mice at day nine after MOG immunization (**e**), or after in vitro stimulation with MOG plus IL-12 for four days (**f**). **g**, Plot of odds ratio versus odds ratio for motif-enrichment data from *Rptor*^{Il17aCre} versus WT and 2-DG versus PBS samples. **h**, Flow cytometry of T-bet and TCF-1 expression in YFP$^+$ cells from Il17aCre (R26R^{eYFP}) mice cultured with MOG plus IL-12, and vehicle or 2-DG (1 mM). Data are representative of one (**a–d**, **g**), four (**e**, **f**) or three (**h**) independent experiments.

transcription factor STAT4, but showed only a small defect in IL-23-induced STAT3 phosphorylation (Extended Data Fig. 4j). Altogether these results show that mTORC1 facilitates the transdifferentiation of T_H17 cells into IFN-γ-producing cells with T_H1-like features after antigen stimulation, and that this can be intensified by cytokine signals from IL-12 and, to a lesser extent, IL-23.

Transcriptome and functional enrichment analyses of Raptor-deficient T_H17 cells identified the acetylation and cholesterol-biosynthesis pathways as amongst the most extensively downregulated (Extended Data Fig. 5a). mTORC1 signalling, Myc targets and metabolic pathways were also attenuated (Extended Data Fig. 5b). Moreover, Raptor-deficient and CD27$^+$ (versus CD27$^-$) T_H17 cells shared downregulated metabolic pathways, including Myc and mTORC1 signalling and nucleotide metabolism (Extended Data Fig. 5c, d); upregulated or downregulated expression changes in CD27$^+$ T_H17 cells were positively correlated with those of Raptor-deficient or wild-type cells, respectively (Extended Data Fig. 5e). In further support of mTORC1 signalling as a differentiating feature of CD27$^+$ and CD27$^-$ cells, *Cd27* expression was upregulated in Raptor-deficient cells (Extended Data Fig. 5f).

We next used Ingenuity pathway analysis (IPA) to identify transcription factors that mediate mTORC1 function, and identified downregulation of Myc, sterol regulatory element-binding proteins (SREBPs) and STATs in Raptor-deficient T_H17 cells (Extended Data Fig. 6a). Flow cytometry validated the lower Myc expression in Raptor-deficient cells (Extended Data Fig. 6b). To explore the metabolic dependence of T_H17 fates, we used the Il17aCre and R26R^{eYFP} systems to delete *Myc* (Extended Data Fig. 6c, d) or *Hmgcr*, which encodes a SREBP-dependent rate-limiting enzyme for cholesterol biosynthesis (Extended Data Fig. 6e, f). T_H17 cells from these mice largely phenocopied Raptor-deficient cells, as exemplified by reduced *Tbx21* and *Il12rb2* expression (Extended Data Fig. 6c, e) and an impaired ability to transdifferentiate into IFN-γ$^+$ cells (Extended Data Fig. 6d, f). Moreover, deletion of *Myc*, *Hmgcr* or *Rptor* all upregulated expression of CD27 on T_H17 cells (Extended Data Fig. 6g). Finally, to gain insight into pathways

© 2019 Springer Nature Limited. All rights reserved.

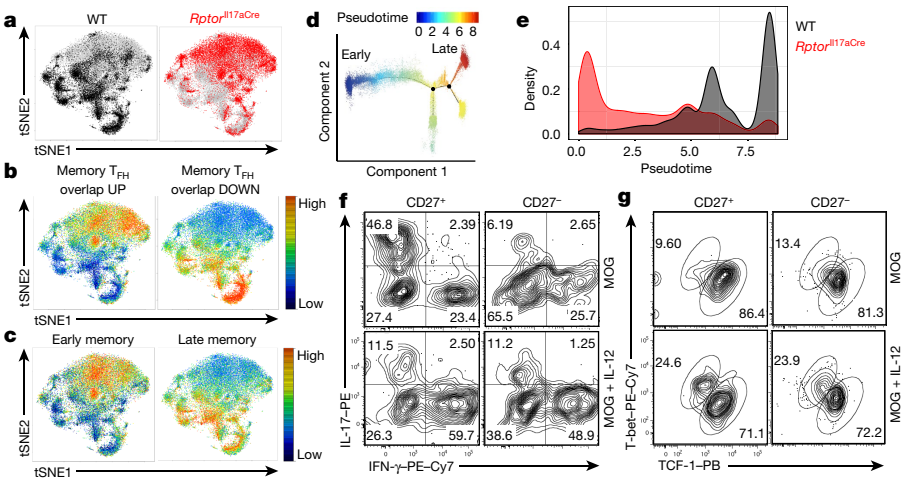


Fig. 4 | Single-cell transcriptomics of T_H17 cells and mTORC1-dependent licensing for T_H1 transdifferentiation. a, Two-dimensional *t*-distributed stochastic neighbour embedding (tSNE) projection of single-cell transcriptomics data from WT (black) and *Rptor*^{I117aCre} (red) YFP$^+$ cells, showing the distribution of individual cells (each dot in the plot represents a single cell; $n = 3$ per genotype; 27,619 total cells). **b**, tSNE visualization of gene signatures of upregulated (UP) and downregulated (DOWN) genes shared between memory CD8$^+$ and T_{FH} cells[13]. **c**, tSNE visualization of 'early memory' and 'late memory' gene signatures[18,19]. **d**, Pseudotime trajectory of single-cell transcriptomics data coloured by pseudotime (arbitrary units; blue and red indicate early and late time, respectively). **e**, Pseudotime trajectory of cellular density. **f**, **g**, Cytokine production (**f**) and T-bet and TCF-1 expression (**g**) by CD27$^+$ and CD27$^-$ YFP$^+$ cells cultured with MOG or with MOG plus IL-12 for four days. Data are representative of one (**a**–**e**) or three (**f**, **g**) independent experiments.

downstream of mTORC1 signalling, we used inhibitors of the S6K and 4E-BP branches. We found that drugs targeting 4E-BP but not S6K signalling impaired the transition of T_H17 cells into IFN-γ-producing cells, although both pathways contributed to cell proliferation (Extended Data Fig. 6h, i). Altogether, our transcriptome and functional perturbation studies indicate that interplay between mTORC1 signalling, especially the 4E-BP branch, and metabolic reprogramming controls T_H17 cell plasticity.

Consistent with downregulation of acetylation in Raptor-deficient cells (Extended Data Fig. 5a), we found impaired binding of acetyl histone H3 to the *Ifng* (IFN-γ gene) promoter (Extended Data Fig. 7a). Also, ATAC-seq (assay for transposase accessible chromatin using sequencing[15]) (Fig. 3a and Extended Data Fig. 7b, c) showed that Raptor-deficient cells had less accessibility in the *Ifng* promoter, corresponding to T-bet-binding sites (Fig. 3b), and in the promoters for *Tbx21* and *Il12rb2* (Extended Data Fig. 7d). Searches for transcription-factor binding motifs in accessible regions of ATAC-seq reads revealed enriched occupancy of factors from the TCF–LEF family in Raptor-deficient cells (Fig. 3c, d, Extended Data Fig. 7e, f). Next we carried out in-depth footprint analysis, in which we superimposed our ATAC-seq results with publicly available TCF-1 chromatin immunoprecipitation (ChIP)-seq data sets. We identified TCF–LEF-binding motifs in *Rptor*^{I117aCre} but not in wild-type open-chromatin regions (Extended Data Fig. 8a); by contrast, we found many more T-bet-binding motifs in the wild-type open-chromatin regions (Extended Data Fig. 8a). To validate the observation of enhanced TCF-1 binding to target genes in Raptor-deficient T_H17 cells, we chose two candidate genes—*Il6ra* and *Lrig1*[16,17] (Extended Data Fig. 8b)—for ChIP assays, and found much greater binding of TCF-1 to these genes in Raptor-deficient cells than in wild-type cells (Extended Data Fig. 8c). Moreover, co-staining of TCF-1 and T-bet in T_H17 cells identified two reciprocal populations, with an increased TCF-1^{hi} T-bet^{lo} but a reduced TCF-1^{lo} T-bet^{hi} population upon loss of Raptor (Fig. 3e). Following culture with MOG and IL-12, wild-type cells became almost exclusively TCF-1^{lo} T-bet^{hi}, while Raptor deficiency reduced this transition (Fig. 3f). Similarly, Myc deletion increased the TCF-1^{hi} T-bet^{lo} population at the expense of the TCF-1^{lo} T-bet^{hi} subset (Extended Data Fig. 8d). Altogether, TCF–LEF factors are enriched in the genomic landscape of Raptor-deficient T_H17 cells, and reciprocal expression of T-bet and TCF-1 in T_H17 cells depends upon mTORC1.

To directly assess the involvement of metabolism, we performed ATAC-seq analysis of T_H17 cells isolated from MOG-immunized mice treated with 2-DG (Fig. 3a). Remarkably, 2-DG-treated samples, like Raptor-deficient cells, showed reduced chromatin accessibility at the *Ifng*, *Tbx21* and *Il12rb2* promoters (Fig. 3b and Extended Data Fig. 7d), and enriched occupancy of TCF–LEF factors (Fig. 3g and Extended Data Fig. 8e). Furthermore, 2-DG treatment impaired the generation of the TCF-1^{lo} T-bet^{hi} population (Fig. 3h) and IFN-γ expression in T_H17 cells in vitro and in vivo (Extended Data Fig. 8f, g). These results support the metabolic dependence of chromatin accessibility and reciprocal T-bet and TCF-1 expression in T_H17 cells.

We next used single-cell RNA sequencing (scRNA-seq) to dissect cellular heterogeneity in an unbiased manner. Wild-type and Raptor-deficient T_H17 cells had largely distinct distribution patterns (Fig. 4a and Extended Data Fig. 9a–c). Functional enrichment analysis on the differentially expressed genes within each cluster revealed the underlying cellular processes associated with each subpopulation (Extended Data Fig. 9d). Clusters composed mainly of Raptor-deficient cells were enriched with memory-associated gene signatures[1–3,13], while wild-type-dominant clusters were reciprocally enriched with genes downregulated in memory cells (Fig. 4b and Extended Data Fig. 9e). To explore potential differences in temporal activation, we superimposed our scRNA-seq data with published data sets from CD8$^+$ memory T cells stimulated once versus multiple times[18], respectively designated as 'early memory' (less mature) and 'late memory' (terminal differentiation) signatures[19]. Raptor-deficient and wild-type cells were enriched with 'early memory' and 'late memory' gene signatures, respectively (Fig. 4c and Extended Data Fig. 9e). Moreover, enrichments for T-bet target genes and glycolysis were mainly observed in wild-type-dominant clusters (Extended Data Fig. 9e, f), whereas *Rptor*^{I117aCre}-dominant clusters showed increased expression of memory-associated genes *Bcl2*, *Cd27* and *Tcf7* (refs [9–11]; Extended Data Fig. 9g). Therefore, single-cell transcriptomics reveals a marked heterogeneity of T_H17 cells, with reciprocal expression of stemness-like and terminal-differentiation signatures, and highlights mTORC1-dependent shaping of these heterogeneous features.

To reconstruct developmental trajectory, we analysed our scRNA-seq data using Monocle 2 for pseudotemporal ordering of cells (pseudotime)[20]. Using a group of highly expressed and dispersed genes (Extended Data Fig. 10a), we derived a relative trajectory with

© 2019 Springer Nature Limited. All rights reserved.

pronounced differences in the pseudotime assignment of the nine clusters (Fig. 4d and Extended Data Fig. 10b). Projection of gene signatures onto the pseudotime trajectory revealed the assignment of 'early memory' and 'late memory' signatures with early and late pseudotime, respectively, corroborating our analysis (Extended Data Fig. 10c). The 'T-bet targets' signature (Extended Data Fig. 10c) and *Tbx21* and *Ifng* expression (Extended Data Fig. 10d, e) were restricted to branches of late pseudotime, suggesting late acquisition of T_H1 features. In addition, Raptor-deficient T_H17 cells had an increased propensity for early pseudotime assignment (Fig. 4e and Extended Data Fig. 10f), suggesting altered pseudotemporal ordering. Moreover, *Cd27* and *Tcf7* were expressed primarily in early pseudotime and associated with $Rptor^{Il17aCre}$-dominant clusters (Extended Data Fig. 10g, h). These analyses suggest a role for mTORC1 in promoting the terminal differentiation of T_H1-like cells from $CD27^+$ T_H17 cells.

Although freshly isolated $CD27^+$ T_H17 cells expressed low levels of IFN-γ and T-bet (Fig. 1b and Extended Data Fig. 10i), they acquired the ability to express these proteins to a similar extent as $CD27^-$ cells after antigen stimulation (Fig. 4f, g), indicating a much greater induction than in $CD27^-$ cells (Extended Data Fig. 10j). To further examine the role of mTORC1 in this process, we isolated $CD27^+$ or $CD27^-$ YFP^+ cells from wild-type or $Rptor^{Il17aCre}$ mice and transferred them into recipients, which were subsequently immunized with MOG. Although a substantial number of wild-type $CD27^+$ cells became $CD27^-$, most Raptor-deficient cells retained their CD27 expression (Extended Data Fig. 10k), indicating the arrested differentiation of $CD27^+$ into $CD27^-$ cells. Moreover, Raptor-deficient $CD27^+$ cells were less efficient at transitioning into $CD27^-$ cells after in vitro MOG and IL-12 stimulation (Extended Data Fig. 10l). Altogether, these functional data further support our hypothesis that mTORC1 drives the transition of $CD27^+$ cells to $CD27^-$ cells.

Our findings suggest that T_H17 responses in chronic autoimmune disease are phenotypically, transcriptionally and metabolically heterogeneous, encompassing a $CD27^+$ TCF-1^{hi} subset with inferred stemness features and low anabolic metabolism, and a reciprocal $CD27^-$ T-bet^{hi} subset (Extended Data Fig. 10m). The $CD27^+$ TCF-1^{hi} subset can develop into the terminally differentiated $CD27^-$ T-bet^{hi} subpopulation with robust IFN-γ expression, but this transition is arrested upon mTORC1 deletion or metabolic perturbation. These results highlight that metabolic heterogeneity of T_H17 cells underlies lineage stability and plasticity, and point to a causative effect of metabolic reprogramming on these late developmental processes. Our results suggest that TCF-1 expression is sensitive to altered anabolic metabolism and may serve as a metabolic gatekeeper to preserve lineage identity and the stability of effector T cell lineages. From a therapeutic perspective, much as targeting cancer stem cells is an important goal for successful anticancer therapy, the identification of stem-like T_H17 cells opens up new avenues for therapeutic intervention in chronic inflammatory conditions.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0806-7.

Received: 30 November 2017; Accepted: 30 October 2018;
Published online 19 December 2018.

1. Im, S. J. et al. Defining $CD8^+$ T cells that provide the proliferative burst after PD-1 therapy. *Nature* **537**, 417–421 (2016).
2. He, R. et al. Follicular CXCR5-expressing $CD8^+$ T cells curtail chronic viral infection. *Nature* **537**, 412–416 (2016); erratum **540**, 470 (2016).
3. Leong, Y. A. et al. $CXCR5^+$ follicular cytotoxic T cells control viral infection in B cell follicles. *Nat. Immunol.* **17**, 1187–1196 (2016).
4. Gagliani, N. et al. Th17 cells transdifferentiate into regulatory T cells during resolution of inflammation. *Nature* **523**, 221–225 (2015).
5. Lee, Y. K. et al. Late developmental plasticity in the T helper 17 lineage. *Immunity* **30**, 92–107 (2009).
6. Hirota, K. et al. Fate mapping of IL-17-producing T cells in inflammatory responses. *Nat. Immunol.* **12**, 255–263 (2011).
7. Hirota, K. et al. Plasticity of Th17 cells in Peyer's patches is responsible for the induction of T cell-dependent IgA responses. *Nat. Immunol.* **14**, 372–379 (2013).
8. Gaublomme, J. T. et al. Single-cell genomics unveils critical regulators of Th17 cell pathogenicity. *Cell* **163**, 1400–1412 (2015).
9. Hendriks, J. et al. CD27 is required for generation and long-term maintenance of T cell immunity. *Nat. Immunol.* **1**, 433–440 (2000).
10. Zhou, X. et al. Differentiation and persistence of memory $CD8^+$ T cells depend on T cell factor 1. *Immunity* **33**, 229–240 (2010).
11. Grayson, J. M., Zajac, A. J., Altman, J. D. & Ahmed, R. Cutting edge: increased expression of Bcl-2 in antigen-specific memory CD8+ T cells. *J. Immunol.* **164**, 3950–3954 (2000).
12. Kalia, V. et al. Prolonged interleukin-2Rα expression on virus-specific $CD8^+$ T cells favors terminal-effector differentiation *in vivo*. *Immunity* **32**, 91–103 (2010).
13. Choi, Y. S. et al. Bcl6 expressing follicular helper CD4 T cells are fate committed early and have the capacity to form memory. *J. Immunol.* **190**, 4014–4026 (2013).
14. Buck, M. D., Sowell, R. T., Kaech, S. M. & Pearce, E. L. Metabolic instruction of immunity. *Cell* **169**, 570–586 (2017).
15. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
16. Choi, Y. S. et al. LEF-1 and TCF-1 orchestrate T_{FH} differentiation by regulating differentiation circuits upstream of the transcriptional repressor Bcl6. *Nat. Immunol.* **16**, 980–990 (2015).
17. Xu, L. et al. The transcription factor TCF-1 initiates the differentiation of T_{FH} cells during acute viral infection. *Nat. Immunol.* **16**, 991–999 (2015).
18. Wirth, T. C. et al. Repetitive antigen stimulation induces stepwise transcriptome diversification but preserves a core signature of memory $CD8^+$ T cell differentiation. *Immunity* **33**, 128–140 (2010).
19. Muranski, P. et al. Th17 cells are long lived and retain a stem cell-like molecular signature. *Immunity* **35**, 972–985 (2011).
20. Qiu, X. et al. Single-cell mRNA quantification and differential analysis with Census. *Nat. Methods* **14**, 309–315 (2017).

Acknowledgements We acknowledge M. Hendren and A. KC for animal work; the Immunology FACS core facility at St Jude Children's Research Hospital for cell sorting; and N. Chapman and Y. Wang for editing of the manuscript. This work was supported by the National Institutes of Health (NIH; grants AI105887, AI101407, AI131703, CA176624 and NS064599 to H.C.; CA021765 to B.X. and Y.F.) and the National Multiple Sclerosis Society (to H.C.).

Author contributions P.W.F.K. conceived, designed and performed in vitro and in vivo experiments and bioinformatic analyses, analysed data, and wrote the manuscript. X.C. performed scRNA-seq analysis and developed the latent cellular state analysis (LCA) algorithm. S.A.L., A.A.H. and T.-L.M.N. helped to perform immunological experiments. B.X., Y.C. and Y.F. helped with ATAC-seq analysis. Y.D. and G.N. helped with microarray and scRNA-seq analyses. S.R. performed ChIP assays. W.C., C.R. and J.E. helped with scRNA-seq analyses. K.Y. performed early in vivo studies. P.V. did histopathology analysis. H.C. helped to conceive and design experiments, co-wrote the manuscript, and provided overall direction.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0806-7.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0806-7.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to H.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

Mice. Mice were housed and bred at the St Jude Children's Research Hospital animal-care facilities in specific pathogen-free conditions. C57BL/6, *Tcra*$^{-/-}$, CD45.1^{+}, Il17aCre, *Rag1*$^{-/-}$ and 2D2-transgenic mice were purchased from the Jackson Laboratory. *Rptor*^{fl}, *Myc*^{fl} and *Hmgcr*^{fl} mice were as described[21–23]. Cre-expressing mice were used as controls, and littermates were used when possible. Mice were backcrossed for at least ten generations to the C57BL/6 background strain. Female and male mice were used at 6–10 weeks of age. Sample sizes were selected to maximize the chance of uncovering mean differences that were also statistically significant. Age- and sex-matched mice were assigned randomly to experimental and control groups. The assessment of EAE scores and histopathology examination was performed in a blinded fashion. For bone-marrow chimaeras, *Tcra*$^{-/-}$ mice were sublethally irradiated for a total of 500 Rads before receiving three million CD90.2-depleted bone-marrow donor cells. Mice remained on antibiotic (Baytril) water for 2–3 weeks, and, after 6 weeks, reconstitution was determined by flow cytometry analysis of blood samples. Mice were used 6–8 weeks after chimaera generation for experiments. All experiments with mice were conducted in accordance with the St Jude Children's Research Hospital institutional policies, and animal protocols were approved by the Institutional Animal Care and Use Committee of St Jude Children's Research Hospital.

EAE induction. Mice were immunized subcutaneously at four injection sites with a total of 200 μl of emulsified incomplete Freund's adjuvant supplemented with 1 mg *Mycobacterium tuberculosis* strain H37Ra (Difco) (complete Freund's adjuvant; CFA) and 200 μg myelin oligodendrocyte glycoprotein (amino acids 35–55; $MOG_{35–55}$), and received intraperitoneal injections of 200 ng pertussis toxin (PTX; List Biological Laboratories) at the time of immunization and 48 h later. Mice were observed daily for clinical signs and scored as follows: normal mouse; 0, no overt signs of disease; 1, limp tail; 2, limp tail plus hindlimb weakness; 3, total hindlimb paralysis; 4, hindlimb paralysis plus 75% of body paralysis (forelimb paralysis/weakness); 5, moribund.

FACS, antibodies and reagents. For analysis of surface markers, cells were stained in PBS (Gibco) containing 2% (*w*/*v*) bovine serum albumin (BSA; Sigma). Surface proteins were stained for 30 min on ice. CellTrace Violet labelling was performed according to the manufacturer's instructions (Thermo Fisher). Intracellular staining was performed on sorted YFP^{+} cells with the Foxp3/transcription factor staining buffer set according to the manufacturer's instructions (eBioscience). Intracellular staining for cytokines was performed with a fixation/permeabilization kit (BD Biosciences). Caspase-3 staining was performed using instructions and reagents from the 'active caspase-3 apoptosis kit' (BD Biosciences). We used 7-aminoactinomycin D (7-AAD; Sigma) for dead-cell exclusion at 2.5 μg ml^{-1}. The following antibodies were used: anti-CD27 (LG.7F9), anti-CD45.2 (104), anti-IFN-γ (XMG1.2), anti-KLRG1 (2F1), anti-T-bet (eBio4B10) (all from eBioscience); anti-CD43 (activation glycoform; 1B11), anti-CD44 (IM7), anti-CD45.1 (A20), anti-CD62L (MEL-14), anti-CD127 (A7R34), anti-CCR6 (29-2L17), anti-CXCR3 (CXCR3-173), anti-IL-17 (TC11-18H10.1), anti-Ly6C (HK1.4), anti-PD-1 (29F.1A12), anti-Sca-1 (D7), anti-TCR-β (H57-597) (all from Biolegend); anti-CD95 (Jo2), anti-Foxp3 (FJK-16 s), anti-ROR-γt (Q31-378), anti-pSTAT3 (pY705), anti-pSTAT4 (pY693) (all from BD Biosciences); and anti-CD4 (RM4-5) (from SONY), anti-p4E-BP1 T37/46 (236B4), anti-pS6 S235/236 (D57.2.2E), and anti-TCF-1 (C63D9) (from Cell Signaling Technology). Flow cytometry data were analysed using Flowjo 9.7.7 (Tree Star). Cytokines IL-12 (BD Biosciences) and IL-23 (R&D) were resuspended at 5 μg ml^{-1} and 20 μg ml^{-1}, respectively in 0.5% BSA/Dulbecco's DPBS and used at final concentrations as indicated.

T-cell culture after immunization. Mice were immunized with $MOG_{35–55}$ emulsified with CFA as described above, without PTX injection. Spleen or dLN samples were adjusted to 3 × 10^{6} cells per ml in Click's media containing 10% fetal bovine serum (FBS) and 1% antibiotics and glutamine, and cultured in the presence of MOG (25 μg ml^{-1}) alone, MOG plus IL-12 (5 ng ml^{-1}), or MOG plus IL-23 (20 ng ml^{-1}). In some experiments, CellTrace Violet (Life Technologies) was used according to the manufacturer's instructions in order to label cells to assess cell division. CD27^{+} and CD27^{-} cell fractions within YFP^{+} cells were sorted using a Reflection (i-Cyt) and CD27 antibodies described above. Cells were cultured for three (proliferation) or four (cytokine production) days. To assess cytokine production by FACS, cells were stimulated with phorbol 12-myristate 13-acetate (PMA; 50 ng ml^{-1}; Sigma), ionomycin (0.5 μM, Sigma), and monensin (1/1,000, BD Biosciences) for 4 h at 37 °C.

T_H17 differentiation from naive CD4 T cells. Lymphocytes from spleen and dLN were pooled and sorted using a Reflection (i-Cyt). Sorted naive CD4^{+} T cells (CD4^{+} CD25^{-} CD62L^{+} CD44^{-}) were activated in vitro for four to five days with 2 μg ml^{-1} anti-CD3 (2C11; Bio X Cell), 2 μg ml^{-1} anti-CD28 (37.51; Bio X Cell), human TGF-β1 (2 ng ml^{-1}; R&D) and mouse IL-6 (20 ng ml^{-1}; BD Biosciences). Cytokine expression was assessed after PMA/ionomycin stimulation as described above.

ChIP quantitative polymerase chain reaction (qPCR) assay. ChIP was performed as described[24]. Briefly, cells were cross-linked in 1% formaldehyde for 5 min at room temperature, then quenched with glycine, and pellets were lysed in cell lysis buffer (25 mM HEPES, pH 7.8, 1.5 mM $MgCl_2$, 10 mM KCl, 0.3% NP-40, 1 mM dithiothreitol (DTT)) with a protease-inhibitor tablet (Roche) for 10 min on ice. Nuclei were pelleted and lysed in nuclear lysis buffer (50 mM HEPES, pH 7.9, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% sodium deoxycholate, 0.2% sodium dodecyl sulfate (SDS)) with a protease-inhibitor tablet for 10 min on ice, before sonication into pieces of a maximum of 500 base pairs (bp) using a Diagenode Bioruptor. Sheared chromatin was cleared of debris and incubated with either normal IgG (Cell Signaling Technology; 1/50) or anti-pan acetyl histone H3 (Millipore; 1/50), or anti-TCF-1 (Cell Signaling Technology; 1/50), and blocker (Active Motif) rotating overnight at 4 °C. Chromatin immunoprecipitation and subsequent DNA purification was performed using the ChIP-IT high sensitivity kit (Active Motif) as per the manufacturer's instructions. We used the following qPCR primers, labelled with the DNA dye Sybr green: *Ifng* promoter F-GAGCCCAAGGAGTCGAAAGGAA; *Ifng* promoter R-CTAGGTCAGCCGATGGCAGCTA[25]; *Il6ra* F-TTCAGTAAG ACGGCAGGTGT; *Il6ra* R-TACGTTTGACTGAGTGGCCT; *Lrig1* F-TAGGGCGTGCACTTACTGAA; *Lrig1* R-CTGAACTTGCCCTTGACTGG; negative control primer set (Active Motif catalogue number 71011) and positive control primer set (Active Motif number 71017). Data analysis was performed using the 'percent input' normalization method and displayed as a percentage of the WT control. To identify additional target genes aside from known TCF-1 target genes[16,17], we computationally identified genes whose genomic regions became more accessible in ATAC-seq data as a result of Raptor deletion. Using this subset of genes, we further compared microarray data from Raptor-deficient and WT cells and identified *Lrig1* as a top candidate.

2-DG treatment in vivo. Mice were treated by intraperitoneal injection with 2-DG (2 g per kg body weight, as described[26]) dissolved in PBS, or with PBS vehicle alone, for consecutive days, starting at one day before immunization with MOG and continuing until the day before euthanization.

In vivo cell transfer. CD27^{+} or CD27^{-} fractions of YFP^{+} cells were sorted after enrichment with CD4 beads (Miltenyi) from spleens and lymph nodes of MOG-immunized WT or *Rptor*^{Il17aCre} mice. Sorting was performed on a Reflection (i-Cyt) cytometer, and sort-purified cells were transferred via retroorbital injection into *Rag1*$^{-/-}$ mice. Because of low cell numbers, we used day 1 after transfer as the baseline reading against which to normalize subsequent analyses at days 14–15, as described[27], for donor-cell abundance in the spleen and lymph nodes. Similar transfer experiments were performed using CD45.1^{+} hosts, except that 2D2-transgenic mice (expressing MOG-antigen-specific T-cell receptors[28]; crossed onto the Il17aCre (R26R^{eYFP}) fate-mapping system) were used as donor mice, and CD4^{+} T cell enrichment was performed at the time of analysis, to facilitate flow cytometric detection of the rare population of donor cells.

Statistical analysis for biological experiments. For biological experiment (non-omics) analyses, we calculated *P*-values by two-way ANOVA, Student's *t*-test (two-sided) for parametric data or Mann–Whitney *U*-test (two-sided) for non-parametric data. Differences between groups were considered statistically significant with a *P*-value cut-off of 0.05 or less. Data are represented as means ± s.e.m. Graph Pad Prism (v. 6.0) was used to perform these statistical analyses.

Microarray analysis. Mice were immunized with MOG as described above. TCR-β^{+} YFP^{+} cells from dLN of WT or *Rptor*^{Il17aCre} (R26R^{eYFP}) mice or TCR-β^{+} YFP^{+} CD44^{int} CD27^{+} and TCR-β^{+} YFP^{+} CD44^{hi} CD27^{-} cells from dLN of Il17aCre (R26R^{eYFP}) mice at day nine post-immunization were analysed with the Affymetrix Mouse Gene 2.0 ST GeneChip array[29], and expression signals were summarized with the robust multi-array average algorithm (Affymetrix Expression Console v1.1) or by fitting a linear model implemented in the R package 'limma'[30]. Lists of differentially expressed genes by 0.5 $\log_2$ fold change or more were analysed using Ingenuity pathway analysis (www.ingenuity.com) to identify underlying Ingenuity canonical pathways and upstream transcription regulators. Gene-set-enrichment analysis (GSEA) within 'hallmark' gene sets was performed as described[31]. For GSEA using manually curated gene signatures from public data sets, refer to the section on 'Analysis for scRNA-seq data' below. Upregulated and downregulated genes under each condition were annotated using hallmark and KEGG gene sets (version 6.0 downloaded from MsigDB[31]), and functional enrichment of specific pathways in the gene sets was performed using Fisher's exact test. Fisher's exact *P*-value was corrected for multiple testing using the Benjamini–Hochberg (BH) method. Pathways were deemed significantly enriched at an FDR of less than 5% and enrichment score of 10 or more. To determine the empirical cumulative distribution, we calculated *P*-values by using the Kolmogorov–Smirnov test. Microarray data have been deposited into the National Center for Biotechnology Information (NCBI)'s Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE107521.

© 2019 Springer Nature Limited. All rights reserved.

Preparation of the ATAC-seq library. Mice were immunized with MOG as described above. TCR-β^+ YFP$^+$ cells were sorted from the dLN or spleen at day nine post-immunization, and lysed in 50 μl ATAC-seq lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630 detergent) on ice for 10 min. Resulting nuclei were pelleted at 500*g* for 10 min at 4 °C. Supernatant was carefully removed with a pipette and discarded. The pellet was resuspended in 50 μl transposase reaction mix (25 μl 2 × TD buffer, 22.5 μl nuclease-free water, 2.5 μl transposase) and incubated for 30 min at 37 °C. After the reaction, the DNA was cleaned up using the Qiagen MinElute kit. The barcoding reaction was run using the NEBNext HiFi kit on the basis of the manufacturer's instructions and amplified for five cycles according to ref. [15], using the same primers. Ideal cycle numbers were determined from 5 μl (of 50 μl) from the previous reaction mix using KAPA SYBRFast (Kapa Biosystems) and 20-cycle amplification on an Applied Biosystems 7900HT. Optimal cycles were determined from the linear part of the amplification curve and the remaining 45 μl of PCR reaction were amplified in the same reaction mix using the optimal cycle number.

ATAC-seq analysis. We obtained 2 × 100 bp paired-end reads from all samples, removed any possible Nextera adapter sequences from the 3′ ends of the reads using cutadapt (version 1.9, paired-end mode, default parameter with '–m 6 –O 20'), and aligned them to the mouse genome mm9 (NCBIM37_um from Sanger; ftp://ftp-mouse.sanger.ac.uk/ref/NCBIM37_um.fa) using the Burrows–Wheeler algorithm[32] (version 0.5.9-r26-dev, default parameter); duplicated reads were then marked with Picard (version 1.65 (1160)) and only non-duplicated reads were kept, using samtools (parameter '-q 1 -F 1024' version 0.1.18 (r982:295)). After adjustment using the Tn5 shift (by which reads were offset by +4 bp for the sense strand and −5 bp for the antisense strand), we separated reads into nucleosome-free, mononucleosome, dinucleosome and trinucleosome by fragment size[15], and generated bigwig files by using the centre 80 bp of fragments and scaling them to 30 M nucleosome-free reads. We observed reasonable nucleosome-free peaks and patterns of mononucleosomes, dinucleosomes and trinucleosomes on the Integrated Genomics Viewer[33] (version 2.3.40), and visualized these using heat maps and aggregation plots centred by transcription start sites (TSSs; Extended Data Fig. 7b). All samples had more than 20 M nucleosome-free reads, and visual inspection indicated adequate data quality (see Extended Data Fig. 7b), so we performed peak-calling for nucleosome-free reads using MACS2[34] (version 2.1.0.20150603, default parameters with '–extsize 200 –nomodel', merged by bedtools[35] if within 100 bp) for individual samples. To ensure replicability, we first finalized nucleosome-free regions for each tissue or genotype, only retaining a called peak if it appeared in more than half of replicates; we then counted nucleosome-free reads from each replicate and drew correlation plots and distributions (Extended Data Fig. 7c). The Pearson correlation coefficients (all greater than 0.95) indicated that our data were highly reproducible across samples. Then we merged finalized nucleosome-free regions from all tissue/genotype iterations, counted nucleosome-free reads for each sample and clustered them by correlation, indicating that the samples separated well by genotype (Fig. 3a and Extended Data Fig. 7c).

To find the differentially accessible regions, we first normalized raw nucleosome-free read counts using the trimmed mean of M-values normalization method, and applied the empirical Bayes statistics test after linear fitting from the voom package[36] (R 3.23, edgeR 3.12.1, limma 3.26.9). We used FDR-adjusted *P*-values of 0.05 as the cut-off for more-accessible regions or less-accessible regions in Raptor-deficient samples. For motif analysis, we further selected 1,000 regions as controls that had *P*-values of more than 0.5 and counts per million (c.p.m.) that were greater than the first quartile of all c.p.m. and least variable according to their median absolute deviation (MAD) score. Finally, we used MAST[37] from the MEME suite[38] (version 4.10.2) to assess scanning-motif matches (in the TRANSFAC database, with Vertebrata only included) in the nucleosome-free regions, and Fisher's exact tests to test whether a motif was significantly correlated ($P < 0.05$) with differentially accessible regions compared with the total number of regions. For footprinting of identified motifs, we first generated bigwig files according to all tags of adjusted reads, and normalized them according to the number of autosome reads to 200 M reads (for example, a sample with 100 M autosome reads would be scaled so as to double the bigwig profile). We then generated average bigwig files from the mean of the replicates at each base pair for each sample, using motif matches within a nucleosome-free region for footprinting and taking the average profile across all motif matches at each base pair from −100 bp from motif-match centres to +100 bp. Finally, the footprinting profiles were smoothed with 10-bp bins.

To analyse the enrichment of TCF–LEF and T-bet binding regions, we selected nucleosome-free differentially accessible regions at $P < 0.05$. We further annotated the peaks as regions that were more accessible in *Rptor*^{Il17aCre} compared with WT samples (KO_Larger), or regions that were less accessible in *Rptor*^{Il17aCre} compared with WT samples (KO_Smaller). TCF-1 ChIP-seq peaks were downloaded from the GEO under accession number GSE52070 and T-bet peaks were downloaded from GSM998272. For each group, differentially accessible peaks were overlapped with TCF-1 or T-bet peaks in order to identify regions that were common between ATAC-seq peaks and ChIP-seq peaks using bedtools (version 2.25.0). Finally, we used FIMO[39] from the MEME suite (version 4.9.0) to scan the overlapping regions with TRANSFAC motifs for LEF–TCF-1 family members or T-bet, and Fisher's exact test to test the significance of enrichment of motifs in the KO_Larger versus KO_Smaller regions described above.

scRNA-seq. Mice were immunized with MOG as described above. TCR-β^+ YFP$^+$ cells from spleen at day nine post-immunization were sorted on an i-Cyt Reflection cell sorter into 15-ml tubes containing complete media, counted, and placed on ice. Some samples were pooled before library construction in order to load sufficient cellular materials into the Chromium Controller instrument (10x Genomics)[40]. The cells were counted and examined for viability using a Luna dual fluorescence cell counter (Logos Biosystems). All samples were spun down at 2,000 r.p.m. for 5 min. The supernatant was removed, and cells were resuspended in 100 μl of 1 × PBS (Thermo Fisher Scientific) plus 0.04% BSA (Amresco). The cells were again counted and checked for viability using a Luna dual fluorescence cell counter (Logos Biosystems). Cell counts ranged from 4×10^5 to 1.5×10^6 cells per millilitre and viability was greater than 98%. Single-cell suspensions were loaded onto the Chromium Controller according to their respective cell counts to generate 6,000 single-cell gel beads in emulsion (GEMs) per sample. Each sample was loaded into a separate channel. Libraries were prepared using the Chromium single cell 3′ v2 library and gel bead kit (10x Genomics). The complementary DNA content of each sample after cDNA amplification of 12 cycles was quantified and quality checked using a high-sensitivity DNA chip with a 2100 Bioanalyzer (Agilent Technologies) to determine the number of PCR amplification cycles to yield sufficient library for sequencing. After library quantification and quality checking by DNA 1000 chip (Agilent Technologies), samples were diluted to 3.5 nM for loading onto the HiSeq 4000 (Illumina) with a 2 × 75 paired-end kit using the following read length: 26 bp read1, 8 bp i7 index, and 98 bp read2. An average of 400,000,000 reads per sample was obtained (approximately 80,000 reads per cell).

Analysis of scRNA-seq data. *Alignment, barcode assignment and unique molecular identifier (UMI) counting.* The Cell Ranger 1.3 Single-Cell software suite (10x Genomics) was implemented to process the raw sequencing data from the Illumina HiSeq run. This pipeline performed demultiplexing, alignment (using the mouse genome mm10 from ENSEMBL GRCm38), and barcode processing to generate gene-cell matrices used for downstream analysis. Specifically, data from three WT and three *Rptor*^{Il17aCre} samples were combined into one data set for consistent filtering, and UMIs mapped to genes encoding ribosomal proteins were removed. Cells with low UMI counts (potentially dead cells with broken membranes) or high UMI counts (potentially two or more cells in a single droplet) were filtered. A small fraction of outlier cells (430) was further removed because of their low transcriptome diversity (meaning that fewer genes were detected than in other cells with a comparable number of captured UMIs). A total of 27,619 cells (WT, 13,295; *Rptor*^{Il17aCre}, 14,324) was captured, with an average of 3,419 messenger RNA molecules (UMIs, median: 2,757; range: 1,500–9,999). We normalized the expression level of each gene to 10,000 UMIs per cell and log-transformed them by adding 0.5 to the expression matrix.

Clustering. We inferred the subpopulation structure of the whole data set by using latent cellular state analysis (LCA)[41], a clustering algorithm developed in house for analysing large-scale scRNA-seq data (C. Cheng et al., manuscript submitted). Briefly, LCA first used singular-value decomposition (SVD) to derive latent cellular states from the expression matrix for individual cells. Significant cellular states were determined using the Tracy–Widom test on eigenvalues. A modified version of spectral clustering was performed on the significant cellular states of individual cells (cellular states were explained by total UMIs ignored) with different numbers of clusters (2–30). The optimal number of clusters was manually selected from top models determined by the silhouette measure for solutions with different numbers of clusters.

Data visualization. Underlying cell variations derived from WT and *Rptor*^{Il17aCre} single-cell gene expression were visualized in a two-dimensional projection by *t*-distributed stochastic neighbour embedding (tSNE). Expression of individual genes or pathway scores was colour-coded (grey, not expressed; from low to high, blue–green–yellow–red) for each cell on tSNE plots.

Generation of gene signatures and statistical analysis for functional enrichment. We used published gene sets that are shared between memory and T_{FH} cells ('memory T_{FH} overlap UP' and 'memory T_{FH} overlap DOWN' for upregulated and downregulated genes, respectively)[13], and early and late CD8$^+$ T cell memory gene sets used previously to show memory features of T_H17 cells ('early memory' and 'late memory')[19]. We generated a *Tbx21* (T-bet)-dependent signature ('T-bet targets') from previously published gene-expression changes in cultured T_H1 cells sufficient or deficient for *Tbx21* ($\log_2$ fold change ≥ 3; GSE38808)[42]. A microarray data set (GSE19825)[12] was analysed using the limma package in R to generate 'CD25^{hi} effector CD8' and 'CD25^{lo} memory-precursor CD8' gene signatures. Specifically, upregulated and downregulated genes in the CD25^{hi} versus CD25^{lo} comparison were ranked after filtering at 5% FDR by $\log_2$ fold change, and 196 upregulated

© 2019 Springer Nature Limited. All rights reserved.

genes (with a $\log_2$ fold change greater than 1) were annotated as '$CD25^{hi}$ effector CD8', and the top 200 downregulated genes (with a $\log_2$ fold change of less than −1) were annotated as '$CD25^{lo}$ memory-precursor CD8'. Similarly, we used a microarray data set (GSE84105)[1] to generate '$CXCR5^+$ exhausted CD8 (Ahmed)' and '$CXCR5^-$ exhausted CD8 (Ahmed)' gene signatures at the same significance level of 5% FDR; total upregulated and downregulated genes were more than 200, so we ranked genes by their $\log_2$ fold change of expression in the $CXCR5^+$ versus $CXCR5^-$ comparison, and used the top 200 upregulated genes as '$CXCR5^+$ exhausted CD8 (Ahmed)' and the top 200 downregulated genes as '$CXCR5^-$ exhausted CD8 (Ahmed)'. Furthermore, we processed RNA-seq data (GSE76279)[2] using the DEseq2 R package version 1.16.1 to generate '$CXCR5^+$ exhausted CD8 (Yu)' and '$CXCR5^-$ exhausted CD8 (Yu)' gene signatures via the same strategy as above. For scRNA-seq analysis, we retained genes from these signatures if they were expressed in at least 10% of cells in any cluster. Gene-signature activity was calculated as the average value for all retained genes. For functional enrichment of genes highly expressed in a specific cluster, we applied the Fisher's exact test (R version 3.3.1) using the specific gene signatures described above, or hallmark, canonical pathways and gene-ontology gene sets (MSigDB[31]).

Pseudotime analysis using Monocle 2. Pseudotime analysis was performed using Monocle 2[20] (v. 2.3.6) with cell ordering based on highly dispersed and highly expressed genes (empirical dispersion/dispersion fit $\geq$ 1.1 and mean expression $\geq$ 0.01; Extended Data Fig. 10a). Resulting data were visualized using functions available in the Monocle 2 package and ggplot2 (v. 2.2.1).

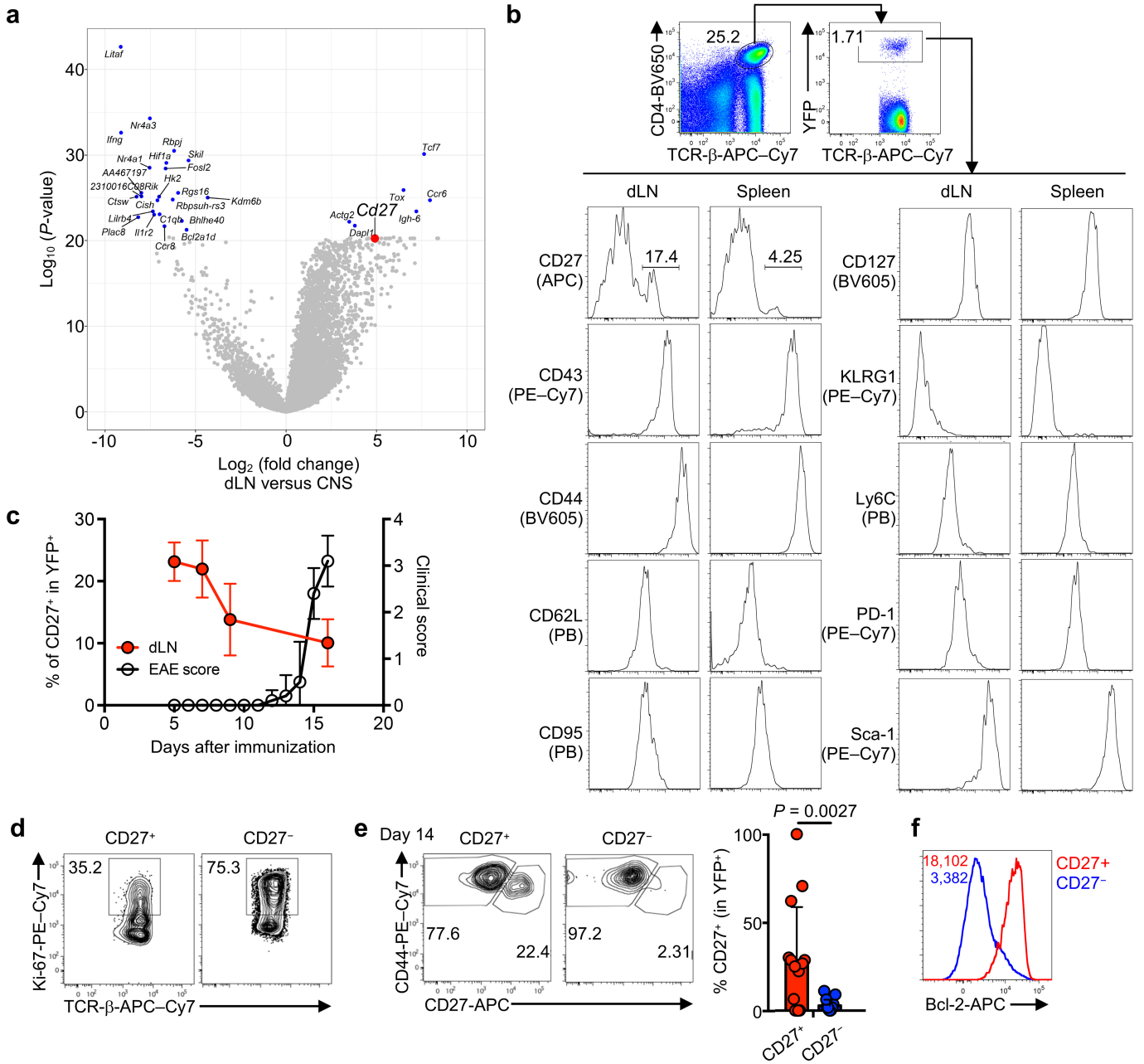
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Microarray data are available via GEO (https://www.ncbi.nlm.nih.gov/geo/) under accession number GSE107521. ATAC-seq and scRNA-seq data are available via GEO under accession number GSE121599.

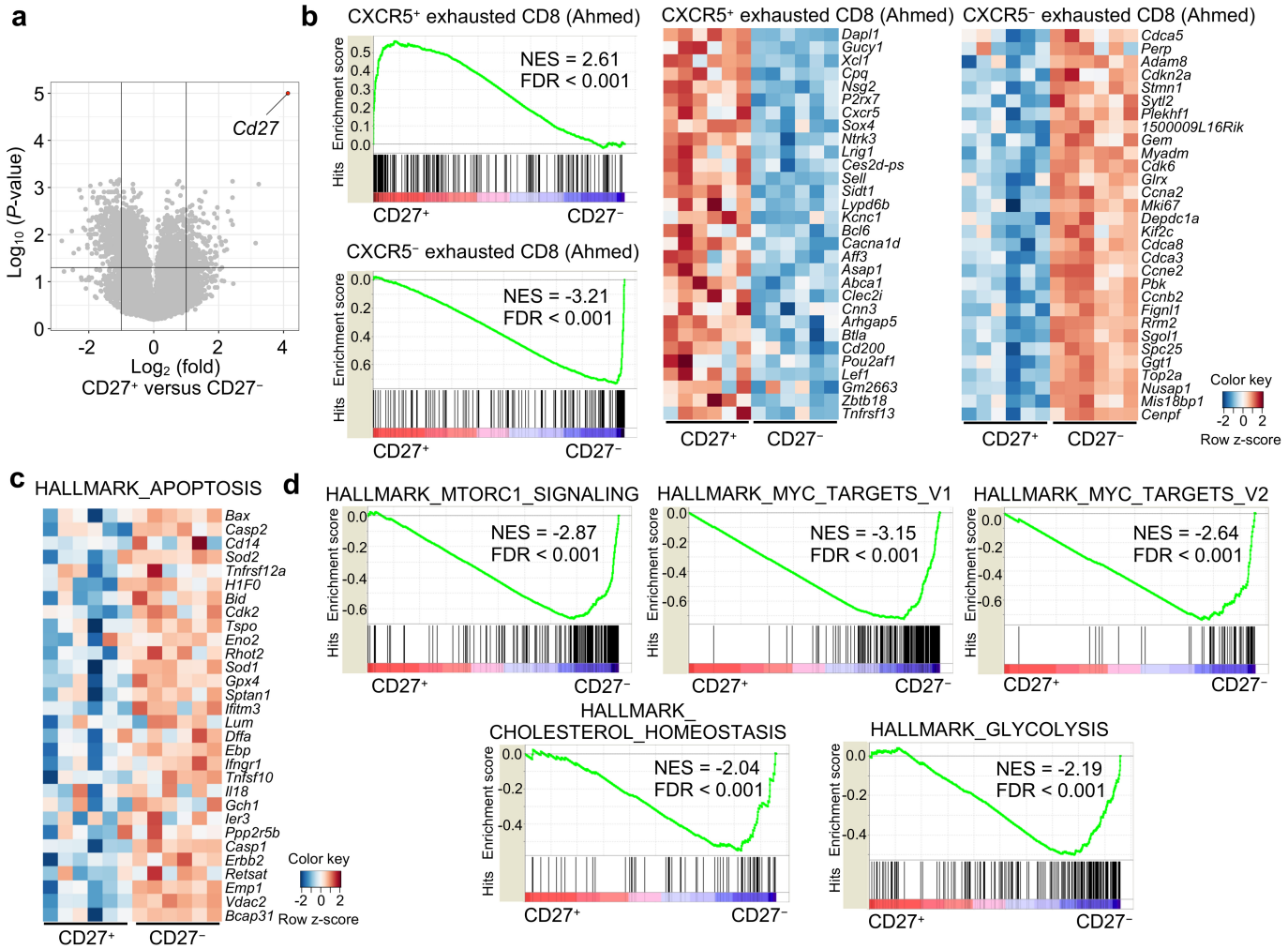
21. Zeng, H. et al. mTORC1 couples immune signals and metabolic programming to establish T_{reg}-cell function. *Nature* **499**, 485–490 (2013).
22. Nagashima, S. et al. Liver-specific deletion of 3-hydroxy-3-methylglutaryl coenzyme A reductase causes hepatic steatosis and death. *Arterioscler. Thromb. Vasc. Biol.* **32**, 1824–1831 (2012).
23. Zeng, H. et al. mTORC1 and mTORC2 kinase signaling and glucose metabolism drive follicular helper T cell differentiation. *Immunity* **45**, 540–554 (2016).
24. Yang, K. et al. Homeostatic control of metabolic and functional fitness of T_{reg} cells by LKB1 signalling. *Nature* **548**, 602–606 (2017).
25. Wang, X. et al. Transcription of Il17 and Il17f is controlled by conserved noncoding sequence 2. *Immunity* **36**, 23–31 (2012).
26. Shi, L. Z. et al. HIF1alpha-dependent glycolytic pathway orchestrates a metabolic checkpoint for the differentiation of TH17 and Treg cells. *J. Exp. Med.* **208**, 1367–1376 (2011).
27. Pepper, M. et al. Different routes of bacterial infection induce long-lived T_H1 memory cells and short-lived T_H17 cells. *Nat. Immunol.* **11**, 83–89 (2010).
28. Bettelli, E. et al. Myelin oligodendrocyte glycoprotein-specific T cell receptor transgenic mice develop spontaneous autoimmune optic neuritis. *J. Exp. Med.* **197**, 1073–1081 (2003).
29. Du, X. et al. Hippo/Mst signalling couples metabolic state and immune function of $CD8\alpha^+$ dendritic cells. *Nature* **558**, 141–145 (2018).
30. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
31. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
33. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
34. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
35. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
36. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
37. Bailey, T. L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* **14**, 48–54 (1998).
38. Bailey, T. L. et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
39. Cuellar-Partida, G. et al. Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics* **28**, 56–62 (2012).
40. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
41. Yang, K. et al. Metabolic signaling directs the reciprocal lineage decisions of $\alpha\beta$ and $\gamma\delta$ T cells. *Sci. Immunol.* **3**, eaas9818 (2018).
42. Zhu, J. et al. The transcription factor T-bet is induced by multiple pathways and prevents an endogenous Th2 cell program during Th1 cell responses. *Immunity* **37**, 660–673 (2012).

© 2019 Springer Nature Limited. All rights reserved.



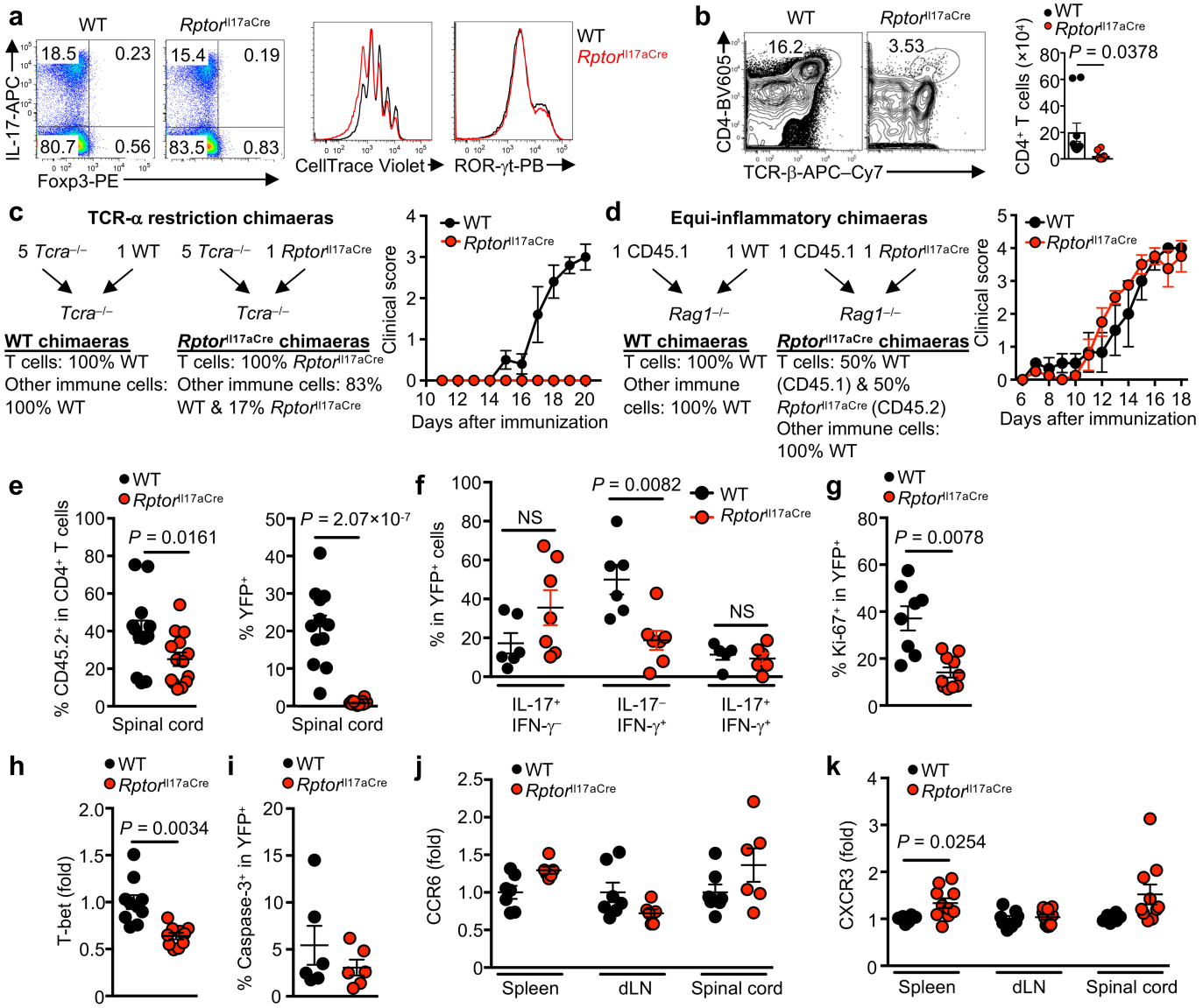
Extended Data Fig. 1 | CD27 expression on T_H17 cells during autoimmunity, and cellular homeostasis of $CD27^+$ and $CD27^-$ T_H17 subsets. a, Analysis of publicly available single-cell transcriptomics data from IL-17–GFP^+ cells in dLN compared with the CNS[8]. **b**, Flow cytometry analysis of putative memory or activation markers on $CD4^+$ TCR-β^+ YFP^+ cells from the dLN and spleen of mice at day nine post MOG-immunization. **c**, Summary of CD27 expression on $CD4^+$ TCR-β^+ YFP^+ cells in dLN (red; $n = 5$, day 5; $n = 3$, day 7; $n = 7$, day 9; $n = 8$, day 16) overlaid with clinical EAE score (black, $n = 5$). **d**, Ki-67 expression in $CD27^+$ and $CD27^-$ cell populations. **e**, Flow cytometry analysis (left) and summary (right) of CD27 expression on transferred $CD27^+$ or $CD27^-$ YFP^+ cells, at day 14 after transfer into $CD45.1^+$ hosts ($n = 14$, $CD27^+$; $n = 15$, $CD27^-$). **f**, Flow cytometry analysis of Bcl-2 expression in $CD27^+$ and $CD27^-$ populations. Data are means ± s.e.m. and representative of three (**b**–**d**, **f**) or at least five (**e**) independent experiments. Numbers in plots represent frequencies of cells in gates; numbers within histograms represent mean fluorescence intensities. The Mann–Whitney U-test (two-sided; non-parametric) was used in **e** to determine statistical significance.

© 2019 Springer Nature Limited. All rights reserved.



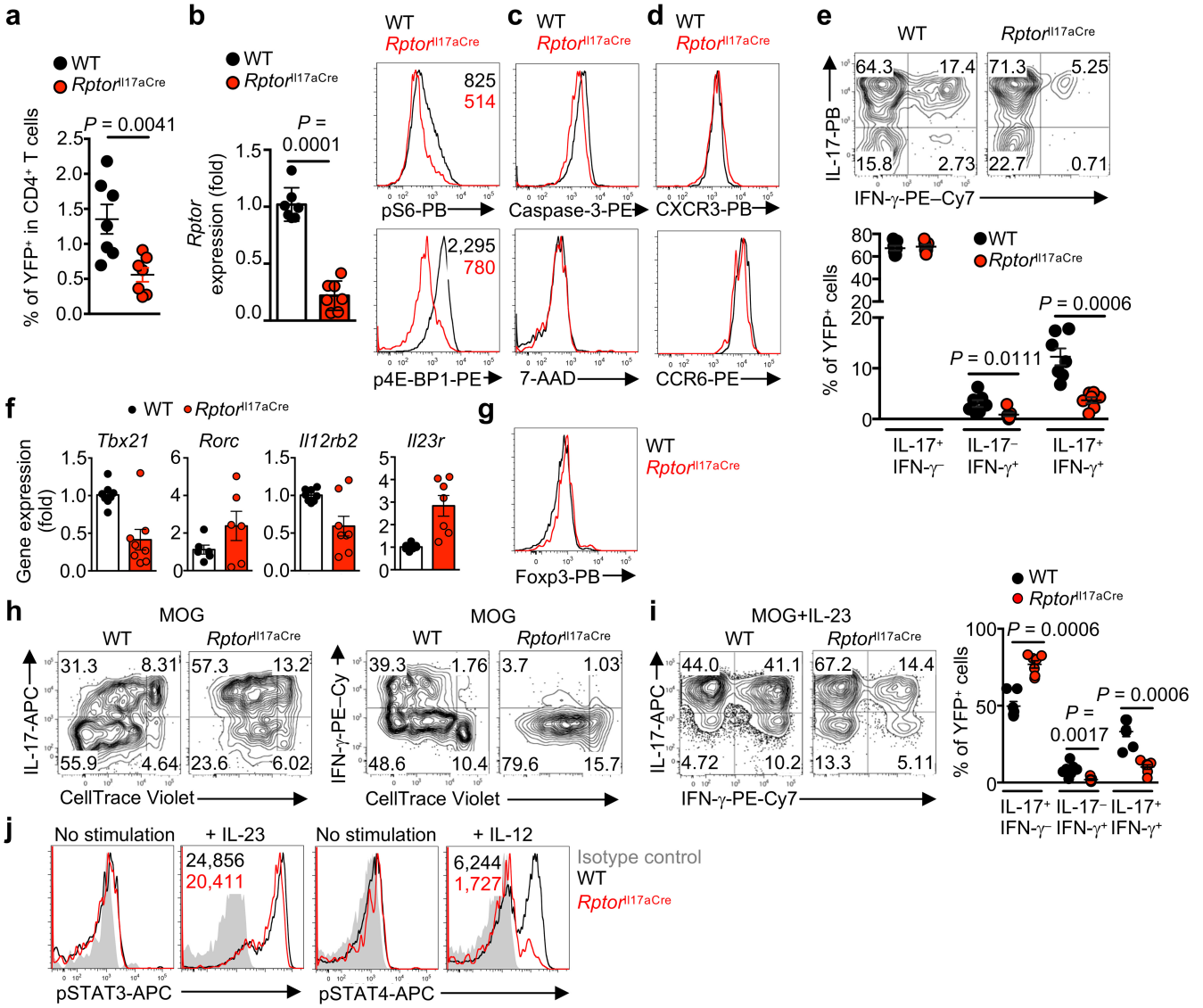
Extended Data Fig. 2 | Gene-expression profiles associated with CD27$^+$ and CD27$^-$ T$_H$17 subsets. a, Volcano plot of transcriptomics data in CD27$^+$ versus CD27$^-$ CD4$^+$ TCR-β$^+$ YFP$^+$ cells. **b**, GSEA plots comparing CD27$^+$ and CD27$^-$ populations using gene sets of antigen-specific CXCR5$^+$ and CXCR5$^-$ exhausted CD8$^+$ T cells from chronic infection. Gene-expression heat maps are normalized by row (*z*-score) for the top 30 leading-edge genes between microarray samples from CD27$^+$ versus CD27$^-$ CD4$^+$ TCR-β$^+$ YFP$^+$ cells, using gene sets derived from CXCR5$^+$ and CXCR5$^-$ exhausted CD8$^+$ T cells[1]. **c**, Gene-expression heat maps normalized by row (*z*-score) for the top 30 leading-edge genes in CD27$^+$ versus CD27$^-$ CD4$^+$ TCR-β$^+$ YFP$^+$ cells, using the apoptosis hallmark gene set. **d**, GSEA plots comparing CD27$^+$ and CD27$^-$ populations using 'hallmark' gene sets, showing the enrichment of mTORC1 signalling, Myc targets, and selective metabolic pathways in CD27$^-$ T$_H$17 cells. Data are from one experiment (**a**–**d**). NES, normalized enrichment score.

© 2019 Springer Nature Limited. All rights reserved.



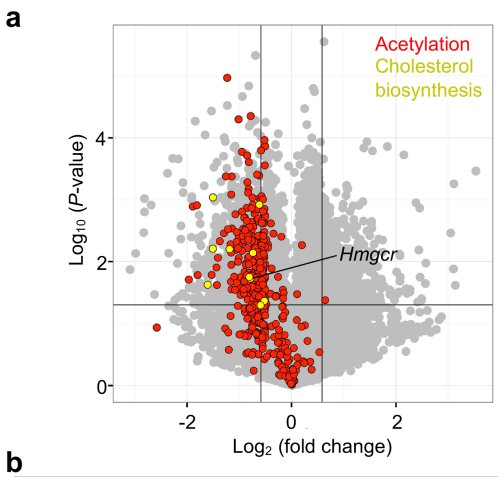
Extended Data Fig. 3 | Cell-intrinsic requirement for Raptor in T_H17 cells. a, Naive CD4$^+$ T cells were differentiated under T_H17-polarizing conditions and analysed for cytokine expression after PMA/ionomycin stimulation in vitro (left), or for proliferation (CellTrace Violet) and ROR-γt expression (right). **b**, Flow cytometry analysis (left) and total number of CD4$^+$ T cells (right) from spinal cord at day 16 post-immunization ($n = 10$, WT; $n = 8$, *Rptor*^{Il17aCre}). **c**, Experimental design for the generation (left) and clinical scores (right) of WT and *Rptor*^{Il17aCre} (R26R^{eYFP}) bone-marrow (BM) chimaeras for restriction of Raptor deficiency specifically to TCR-α-expressing IL-17$^+$ T cells ($n = 5$ per genotype). Specifically, a 5/1 ratio of *Tcra*$^{-/-}$ and *Rptor*^{Il17aCre} (or WT control) BM cells was transferred into sublethally irradiated *Tcra*$^{-/-}$ recipients, followed by reconstitution. In this system, all T cells in the chimaeras were derived from *Rptor*^{Il17aCre} or WT BM cells, whereas most non-T-cell compartments were derived from WT cells. **d**, Experimental design for the generation (left) and clinical scores (right) of WT and *Rptor*^{Il17aCre} (R26R^{eYFP}) BM chimaeras for equal inflammatory conditions. Specifically, mixed BM chimaeras were generated using a 1/1 ratio of congenically marked CD45.2$^+$ *Rptor*^{Il17aCre} (or WT control) and CD45.1$^+$ WT BM-derived cells, which mediated CNS inflammation in EAE ($n = 3$, WT; $n = 4$, *Rptor*^{Il17aCre}). **e**–**k**, The equi-inflammatory chimaeric mice generated in panel **d** were analysed at day 18 post-MOG immunization, for the frequencies of CD4$^+$ T cells positive for CD45.2 or YFP (**e**; $n = 12$, WT; $n = 14$, *Rptor*^{Il17aCre}) and YFP$^+$ CD4$^+$ T cells expressing IL-17 or IFN-γ within the spinal cord (**f**; $n = 6$, WT; $n = 7$, *Rptor*^{Il17aCre}); the expression of Ki-67 (**g**; $n = 8$, WT; $n = 10$, *Rptor*^{Il17aCre}), T-bet (**h**; as fold-change in mean fluorescence intensity after normalization to WT cells) ($n = 10$ per genotype), and active caspase-3 in splenic YFP$^+$ CD4$^+$ T cells (**I**; $n = 6$ per genotype); and the expression of CCR6 (**j**; $n = 7$, WT; $n = 6$, *Rptor*^{Il17aCre}) and CXCR3 (**k**; $n = 8$, WT; $n = 10$, *Rptor*^{Il17aCre}) in YFP$^+$ cells from different organs. Data are means ± s.e.m. and representative of three (**a**, **c**, **i**–**k**) or four (**b**, **d**–**h**) independent experiments. Numbers in plots represent frequencies of cells in gates or quadrants. Student's *t*-test (two-sided) was used in **h** and **k**, and Mann–Whitney U-test (two-sided) was used in **b**, **e**, **f** and **g**, to determine statistical significance.

© 2019 Springer Nature Limited. All rights reserved.



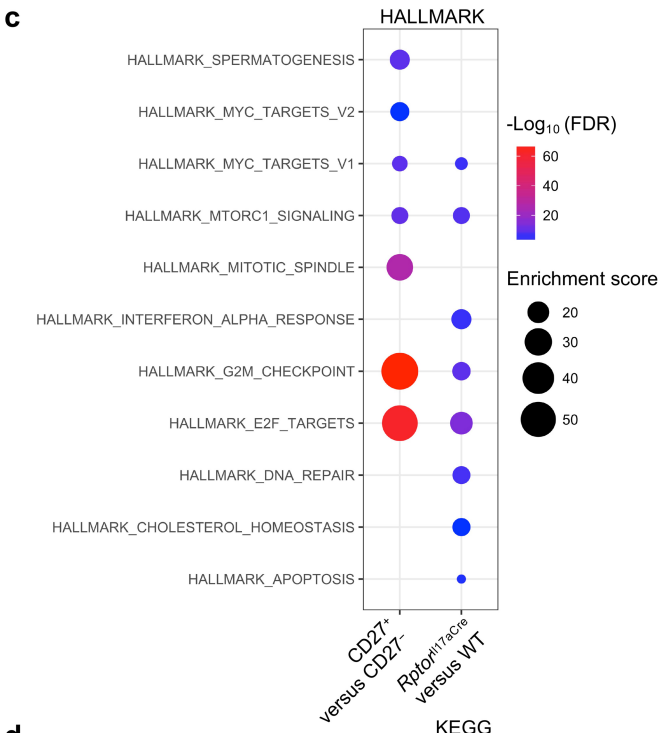
Extended Data Fig. 4 | Raptor deficiency induces selective phenotypic changes in fate-mapped T_H17 cells. a–g, WT and *Rptor*^{Il17aCre} ($R26R^{eYFP}$) mice were immunized with MOG, and YFP$^+$ cells from dLN were analysed at day nine post-immunization. **a**, Frequency of YFP$^+$ cells ($n = 7$ per genotype). **b**, Efficiency of *Rptor* deletion (left; $n = 7$ per genotype) and flow cytometry analysis of phosphorylated S6 (S235/236) and 4E-BP1 (T37/46) (right) in YFP$^+$ cells. **c**, Flow cytometry analysis of active caspase-3 and 7-AAD staining in YFP$^+$ cells. **d**, Flow cytometry analysis of CXCR3 and CCR6 expression on YFP$^+$ cells. **e**, Cytokine production by YFP$^+$ cells from dLN ($n = 7$ per genotype). **f**, Real-time PCR analysis of *Tbx21* ($n = 8$ per genotype), *Rorc* ($n = 6$ per genotype), *Il12rb2* ($n = 7$ per genotype) and *Il23r* ($n = 7$ per genotype) expression in YFP$^+$ cells. **g**, Flow cytometry analysis of Foxp3 expression. **h**, **i**, Cytokine production (**h**, **i**) and proliferation (**h**) of YFP$^+$ cells from dLN of the indicated mice after four days of stimulation with MOG alone (**h**) or with MOG plus IL-23 (**i**) ($n = 7$ per genotype). **j**, Sorted YFP$^+$ cells were stimulated with IL-23 or IL-12 for 30 min in vitro and stained with specific antibodies to phosphorylated STAT3 (left), phosphorylated STAT4 (right), or isotype controls. Data are means ± s.e.m. and representative of seven (**a**), three (**b**–**f**, **j**), two (**g**), or five (**h**, **i**) independent experiments. Numbers in plots represent frequencies of cells in quadrants; numbers within histograms represent mean fluorescence intensities. Student's *t*-test (two-sided) was used in **b**, and Mann–Whitney U-test (two-sided) in **a**, **e**, **i**, to determine statistical significance.

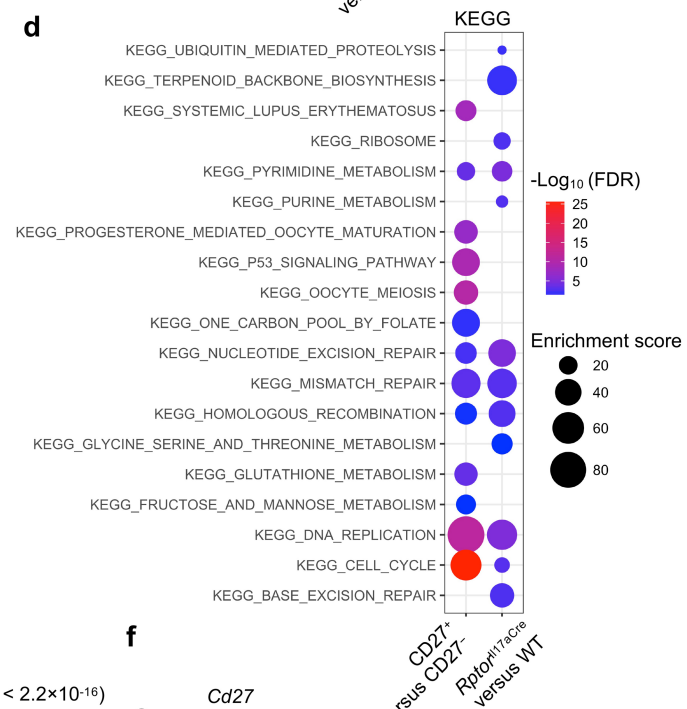
© 2019 Springer Nature Limited. All rights reserved.

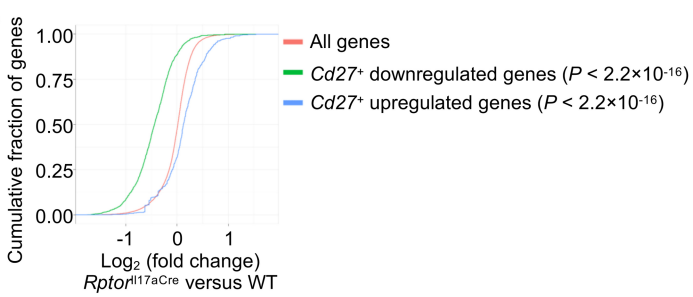


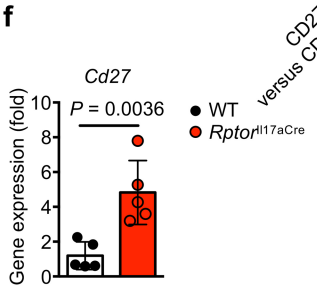
b

Gene sets	NES	FDR q-val
HALLMARK_E2F_TARGETS	-3.0978382	<0.001
HALLMARK_G2M_CHECKPOINT	-2.9223704	<0.001
HALLMARK_MTORC1_SIGNALING	-2.8585272	<0.001
HALLMARK_MYC_TARGETS_V1	-2.6909537	<0.001
HALLMARK_OXIDATIVE_PHOSPHORYLATION	-2.578069	<0.001
HALLMARK_DNA_REPAIR	-2.388436	<0.001
HALLMARK_UNFOLDED_PROTEIN_RESPONSE	-2.367019	<0.001
HALLMARK_INTERFERON_ALPHA_RESPONSE	-2.1688917	<0.001
HALLMARK_MYC_TARGETS_V2	-2.1609168	<0.001
HALLMARK_INTERFERON_GAMMA_RESPONSE	-2.104767	<0.001
HALLMARK_MITOTIC_SPINDLE	-2.0971394	<0.001
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	-1.8384426	6.16E-04
HALLMARK_CHOLESTEROL_HOMEOSTASIS	-1.795519	9.09E-04
HALLMARK_FATTY_ACID_METABOLISM	-1.7904822	9.52E-04
HALLMARK_PI3K_AKT_MTOR_SIGNALING	-1.7734656	0.001114649
HALLMARK_GLYCOLYSIS	-1.7256962	0.0018596
HALLMARK_ANDROGEN_RESPONSE	-1.6734405	0.004365208
HALLMARK_ADIPOGENESIS	-1.644021	0.005527202
HALLMARK_UV_RESPONSE_UP	-1.6272646	0.005887214
HALLMARK_XENOBIOTIC_METABOLISM	-1.5875028	0.008981145
HALLMARK_HEME_METABOLISM	-1.5691196	0.010681462
HALLMARK_PEROXISOME	-1.4813807	0.023629222
HALLMARK_SPERMATOGENESIS	-1.4407699	0.03448933







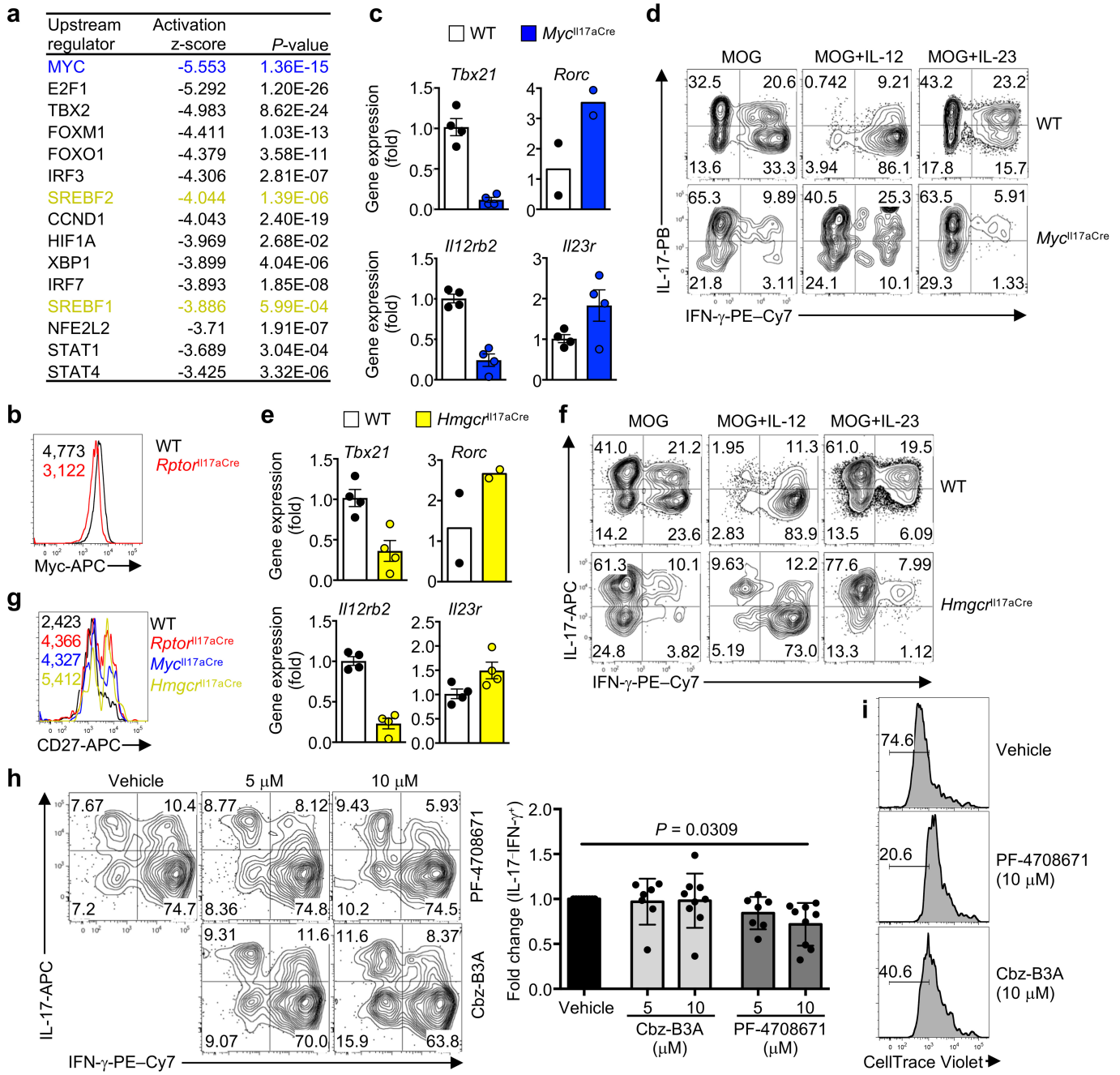


Extended Data Fig. 5 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

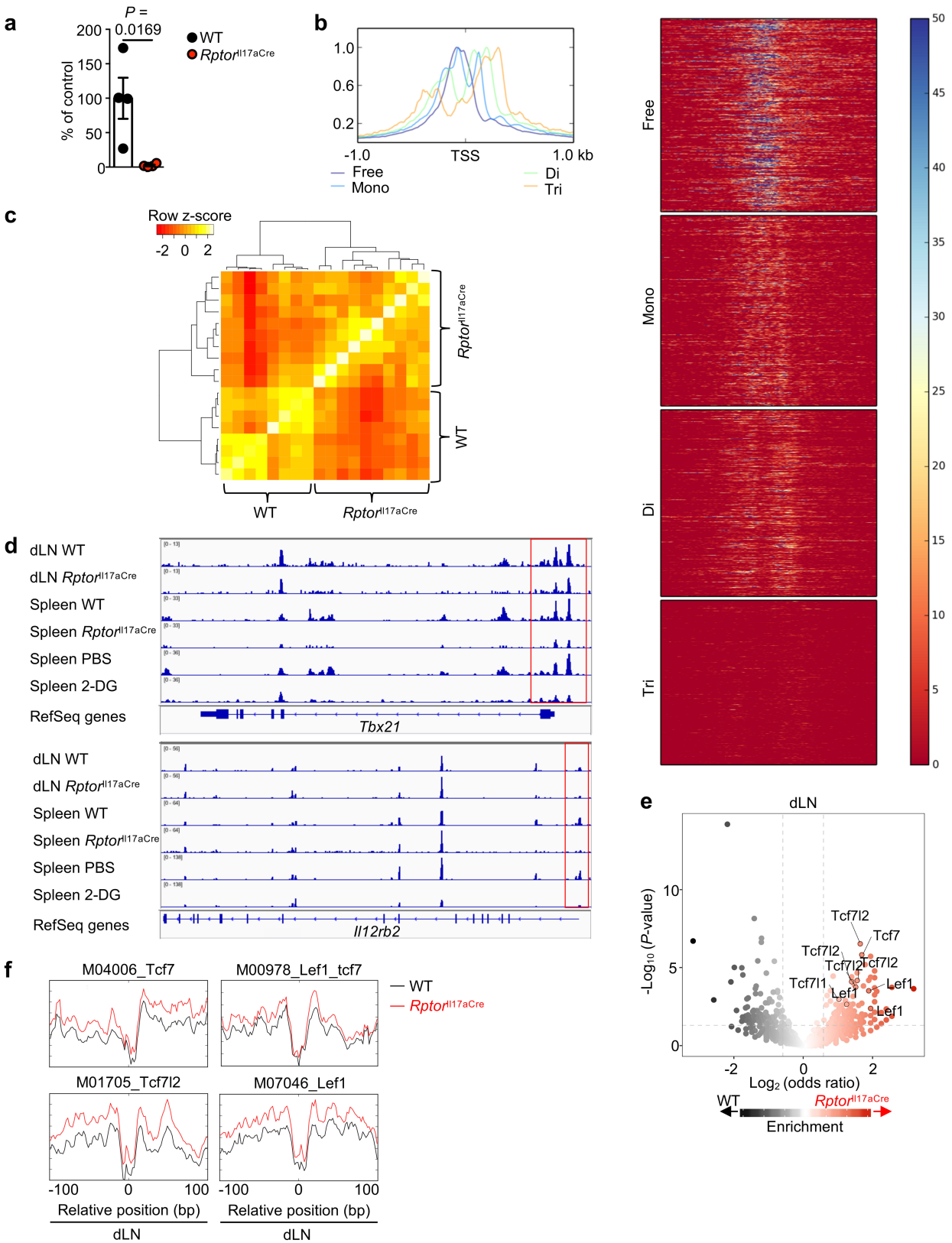
Extended Data Fig. 5 | Altered gene-expression profiles in Raptor-deficient cells, and shared functional pathways with CD27$^+$ T$_H$17 cells. a–d, WT and *Rptor*^{Il17aCre} (R26R^{eYFP}) mice were immunized with MOG, and sorted YFP$^+$ cells from dLN were analysed by microarray. **a**, Expression of individual genes, with vertical lines indicating −1.5-fold and +1.5-fold cut-offs and a horizontal line indicating the $P = 0.05$ cut-off; gene-ontology (GO) gene sets for acetylation and cholesterol biosynthesis are coloured and *Hmgcr* is indicated. **b**, GSEA reveals significant 'hallmark' gene sets that are downregulated in *Rptor*^{Il17aCre} compared with WT cells ($P < 0.05$). **c**, **d**, Comparison of functional enrichment of coregulated gene sets in *Rptor*^{Il17aCre} and CD27$^+$ YFP$^+$ cells. We used the downregulated genes in CD27$^+$ versus CD27$^-$ T$_H$17 cells and in *Rptor*^{Il17aCre} versus WT T$_H$17 cells (with an FDR of less than 0.05 and the top 200 genes, based on fold change) for functional enrichment using hallmark (**c**) and KEGG pathway (**d**) gene sets. Keys indicate FDR values and enrichment scores. **e**, Comparison of microarray analyses of CD27$^+$ versus CD27$^-$ and *Rptor*^{Il17aCre} versus WT samples from mice at day nine post MOG-immunization. Shown are empirical cumulative distribution functions for the changes in expression ($\log_2$ values) of all genes expressed in *Rptor*^{Il17aCre} (R26R^{eYFP}) T$_H$17 cells (red line; changes are relative to expression in WT T$_H$17 cells) and for subsets of genes downregulated (green line) or upregulated (blue line) in CD27$^+$ versus CD27$^-$ T$_H$17 cells (with a FDR of less than 5%). *P*-values were calculated using the Kolmogorov–Smirnov test. **f**, Real-time PCR analysis of *Cd27* expression in WT and *Rptor*^{Il17aCre} YFP$^+$ cells ($n = 5$ per genotype). Data are means ± s.e.m. and from one experiment (**a**–**e**) or representative of two independent experiments (**f**). Student's *t*-test (two-sided) was used in **f** to determine statistical significance.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Anabolic metabolism promotes the transdifferentiation of T_H17 cells into T_H1-like IFN-γ-producing cells. **a**, Ingenuity pathway analysis (IPA) of upstream transcriptional regulators in *Rptor*^{Il17aCre} versus WT samples. **b**, WT and *Rptor*^{Il17aCre} (R26R^{eYFP}) mice were immunized with MOG, and YFP$^+$ cells from dLN were analysed by flow cytometry for intracellular expression of Myc. **c**, **d**, WT and *Myc*^{Il17aCre} (R26R^{eYFP}) mice were immunized with MOG, and YFP$^+$ cells from dLN were analysed by real-time PCR at day nine (**c**; $n = 4$, *Tbx21*; $n = 2$, *Rorc*; $n = 4$, *Il12rb2*; $n = 4$, *Il23r*), or alternatively dLN cells were cultured with MOG, MOG plus IL-12, or MOG plus IL-23 for four days in order to analyse cytokine expression within YFP$^+$ cells (**d**). **e**, **f**, WT and *Hmgcr*^{Il17aCre} (R26R^{eYFP}) mice were immunized with MOG; YFP$^+$ cells were isolated from dLN and analysed by real-time PCR (**e**; $n = 4$, *Tbx21*; $n = 2$, *Rorc*; $n = 4$, *Il12rb2*; $n = 4$, *Il23r*), or dLN cells were cultured with MOG, MOG plus IL-12, or MOG plus IL-23 for four days in order to analyse cytokine expression within YFP$^+$ cells (**f**). **g**, Flow cytometry analysis of CD27 expression on YFP$^+$ cells from WT, *Rptor*^{Il17aCre}, *Myc*^{Il17aCre} and *Hmgcr*^{Il17aCre} mice at day nine post MOG-immunization. **h**, **i**, WT (R26R^{eYFP}) mice were immunized with MOG, and dLN cells were stimulated with MOG and IL-12 in the presence of vehicle, PF-4708671 (an inhibitor of S6K phosphorylation) or Cbz-B3A (an inhibitor of 4E-BP phosphorylation, which in turn suppresses eIF4E-dependent protein translation) at the indicated concentrations for four days to analyse cytokine expression within YFP$^+$ cells (**h**; right, summary plots) ($n = 9$, vehicle; $n = 7$, PF-4708671 (5 μM); $n = 9$, PF-4708671 (10 μM); $n = 7$, Cbz-B3A (5 μM); $n = 9$, Cbz-B3A (10 μM)) and for CellTrace Violet dilution (**i**). Data are means ± s.e.m. and from one experiment (**a**), or representative of four (**b**–**f**, **h**, **i**) or three (**g**) independent experiments. Numbers in plots represent frequencies of cells in gates or quadrants; numbers within histograms represent mean fluorescence intensities. Student's *t*-test (two-sided; parametric) was used to determine statistical significance in panel **h**.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | See next page for caption.

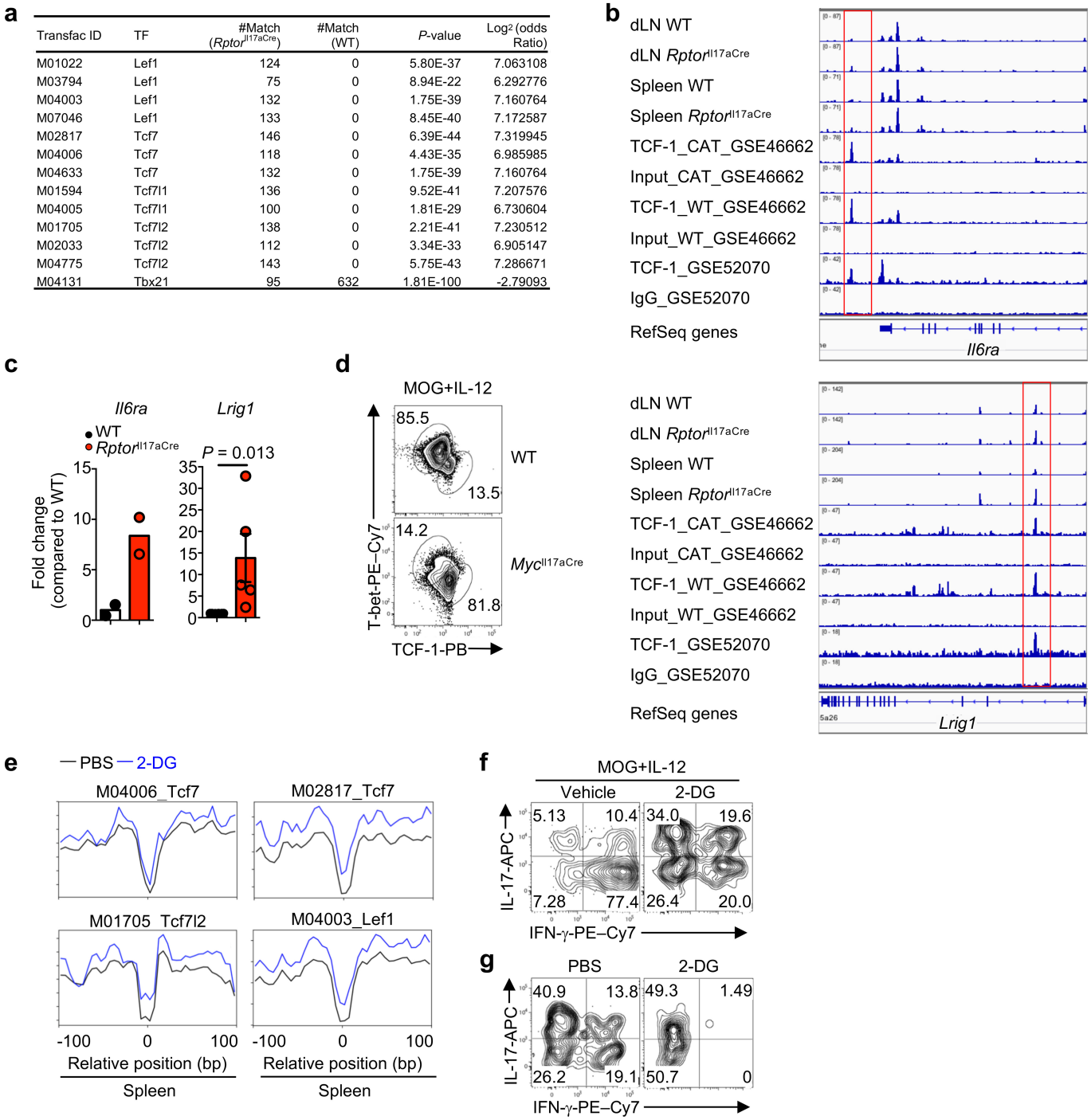
© 2019 Springer Nature Limited. All rights reserved.

Extended Data Fig. 7 | Analysis of histone acetylation, and ATAC-seq overview and specific gene loci. a, ChIP qRT–PCR of pan-acetyl-histone bound to the *Ifng* promoter of WT or *Rptor*^{Il17aCre} (R26R^{eYFP}) YFP^{+} cells from dLN ($n = 4$ per genotype). **b**, **c**, WT and *Rptor*^{Il17aCre} (R26R^{eYFP}) mice were immunized with MOG, and YFP^{+} cells from dLN and spleen at day nine post-immunization were analysed by ATAC-seq. **b**, Density plot and heat maps of a representative individual ATAC-seq sample, demonstrating separation into different fragment lengths indicative of nucleosome-free, mononucleosome, dinucleosome and trinucleosome patterns, consistent with ref. [15]. TSS, transcription start site. **c**, Correlation plot of nucleosome-free fragments. **d**, Nucleosome-free ATAC-seq tracks at the *Tbx21* and *Il12rb2* gene loci, with immediate promoter regions indicated by red boxes. **e**, Summary of ATAC-seq motif-enrichment data, showing $\log_2$ (odds ratio) and $\log_{10}$ (Fisher *P*-value) of cells from dLN. **f**, Tn5 insert sites from ATAC-seq analysis of dLN were aligned to motifs for transcription factors from the TRANSFAC database, and the binding profiles of selected transcription factors of the TCF–LEF family are shown. Data are means ± s.e.m. and representative of three independent experiments (**a**), or from one experiment (**b**–**f**). Student's *t*-test (two-sided; parametric) was used in panel **a** to determine statistical significance.

© 2019 Springer Nature Limited. All rights reserved.

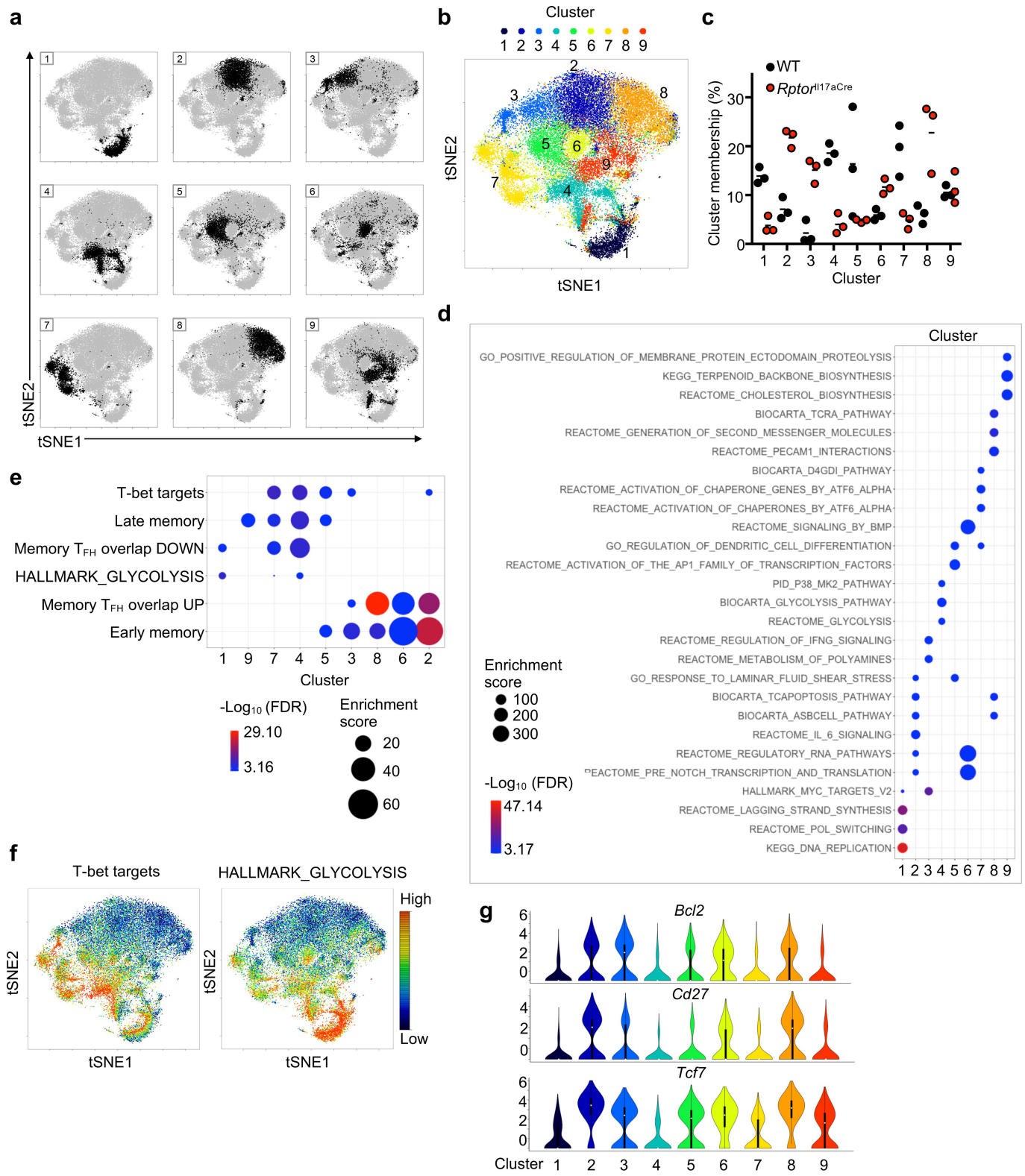
a

Transfac ID	TF	#Match (*Rptor*^{Il17aCre})	#Match (WT)	*P*-value	Log$_2$ (odds Ratio)
M01022	Lef1	124	0	5.80E-37	7.063108
M03794	Lef1	75	0	8.94E-22	6.292776
M04003	Lef1	132	0	1.75E-39	7.160764
M07046	Lef1	133	0	8.45E-40	7.172587
M02817	Tcf7	146	0	6.39E-44	7.319945
M04006	Tcf7	118	0	4.43E-35	6.985985
M04633	Tcf7	132	0	1.75E-39	7.160764
M01594	Tcf7l1	136	0	9.52E-41	7.207576
M04005	Tcf7l1	100	0	1.81E-29	6.730604
M01705	Tcf7l2	138	0	2.21E-41	7.230512
M02033	Tcf7l2	112	0	3.34E-33	6.905147
M04775	Tcf7l2	143	0	5.75E-43	7.286671
M04131	Tbx21	95	632	1.81E-100	-2.79093



Extended Data Fig. 8 | ATAC-seq in-depth analyses, TCF-1 binding activity, and effects of 2-DG on cytokine expression. a, Analysis of common regions in ATAC-seq and ChIP-seq peaks for motifs that bind to TCF–LEF family and T-bet transcription factors, from spleen samples. Numbers of motif matches and associated Fisher's exact test *P*-values and $\log_2$ (odds ratios) are shown; a positive $\log_2$ (odds ratio) value indicates that a motif is more likely to occur in *Rptor*^{Il17aCre} than in WT samples; a negative value indicates that the chance of occurrence is lower in the *Rptor*^{Il17aCre} group. 'E − *x*' denotes '$\times 10^{-x}$'. **b**, Nucleosome-free ATAC-seq tracks at the *Il6ra* and *Lrig1* gene loci, with TCF-1-binding sites indicated by red boxes, based on alignment with TCF-1-binding sites from published data (GEO accession numbers are shown). **c**, ChIP assay to measure TCF-1 binding to *Il6ra* and *Lrig1* gene loci (*Il6ra*, $n = 2$ per genotype; *Lrig1*, $n = 6$ for WT, $n = 5$ for *Rptor*^{Il17aCre}). **d**, Cells from dLN of the indicated mice at day nine post-MOG immunization were cultured for four days with MOG plus IL-12 and sorted on the YFP$^+$ population before intracellular staining. Flow cytometry was used to analyse T-bet and TCF-1 expression in YFP$^+$ cells from WT and *Myc*^{Il17aCre} (R26R^{eYFP}) mice. **e**, Tn5 insert sites from ATAC-seq analysis of YFP$^+$ cells from PBS- or 2-DG-treated mice were aligned to motifs for transcription factors from the TRANSFAC database, and the binding profiles of selected TCF–LEF family transcription factors are shown. **f**, Cytokine expression in dLN YFP$^+$ cells from MOG-immunized Il17aCre (R26R^{eYFP}) mice after culture with MOG and IL-12 for four days in the presence of vehicle (PBS) or 2-DG (1 mM). **g**, Cytokine expression in splenic YFP$^+$ cells from MOG-immunized Il17aCre (R26R^{eYFP}) mice after treatment with 2-DG (2 g per kg of body weight) or PBS. Numbers in plots represent frequencies of cells in gates or quadrants. Data are means ± s.e.m. and from one experiment (**a**, **b**, **e**), or representative of three independent experiments (**c**, **d**, **f**, **g**). Student's *t*-test (two-sided) was used to determine statistical significance in **c**.

© 2019 Springer Nature Limited. All rights reserved.

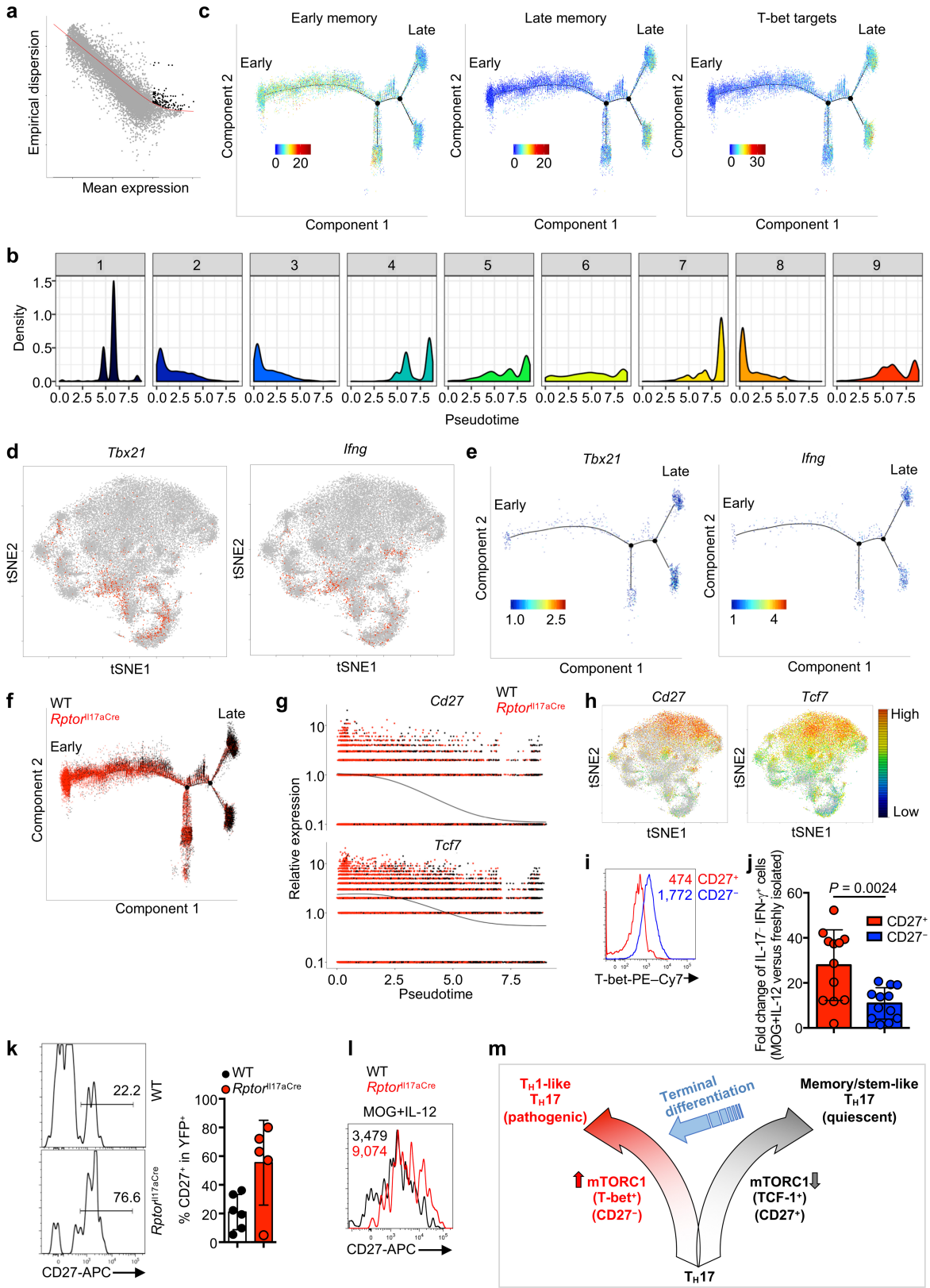


Extended Data Fig. 9 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Fig. 9 | Single-cell transcriptomics analysis. WT and *Rptor*^{Il17aCre} (R26R^{eYFP}) mice were immunized with MOG, and YFP^{+} cells were analysed by single-cell transcriptomics at day nine post-immunization. **a**, Membership of individual cell clusters in two-dimensional tSNE projections from scRNA-seq data. **b**, tSNE visualization of nine clusters partitioned by unsupervised clustering. **c**, Frequencies of WT and *Rptor*^{Il17aCre} cells in different clusters ($n = 3$ per genotype). **d**, Top three enriched gene sets for each of the clusters using 'hallmark', 'canonical' and GO gene sets. For example, genes enriched in cluster 1 by comparison with other clusters were associated with proliferative events. **e**, Summary of cluster-specific functional enrichment analysis via Fisher's exact test, using the signatures of 'T-bet targets', 'late memory', 'memory T_{FH} overlap DOWN', 'HALLMARK_GLYCOLYSIS', 'memory T_{FH} overlap UP' and 'early memory' as described in the Methods. **f**, tSNE visualization of signature scores of 'T-bet targets' and 'HALLMARK_GLYCOLYSIS' expressed in individual cells. **g**, Violin plots of *Bcl2*, *Cd27* and *Tcf7* gene expression amongst the nine clusters. A violin plot combines the box plot and the local density estimation into a single display. The black bars and thin lines within the violin plots indicate, respectively, the interquartile range (first quantile to third quantile) and the entire range of the data (up to 1.5-fold of the interquartile range from first to third quantile); the white dots in the centre indicate the median values. Data are from one experiment (**a**–**g**) ($n = 3$ per genotype).

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Fig. 10 | Pseudotime analysis and experimental validation. a–h, WT and *Rptor*^{Il17aCre} (R26R^{eYFP}) mice were immunized with MOG, and YFP$^+$ cells were analysed by single-cell transcriptomics analysis at day nine post-immunization. **a**, Empirical dispersion and mean expression using the single-cell-analysis toolkit Monocle 2, including the genes used for temporal ordering in black; each grey or black dot represents one gene. The red line shows Monocle's expected dispersion, with more- and less-dispersed genes based on average expression above and below the red line, respectively. **b**, Pseudotime densities for each individual cluster. For example, cluster 1, associated with the proliferative signature, was in the centre of the pseudotime spectrum, while clusters 2, 3 and 8 (early in pseudotime; predominantly Raptor-deficient cells) and clusters 7 and, to a lesser extent, 4 and 5 (late in pseudotime; predominantly WT cells) were on the opposite end of the spectrum. **c**, Projection of signature scores for 'early memory', 'late memory' and 'T-bet targets' onto pseudotime trajectory; the keys indicate the relative scores per cell. **d**, tSNE visualization of *Tbx21* and *Ifng* gene expression. **e**, *Tbx21* and *Ifng* gene expression during pseudotime; cells that did not express *Tbx21* or *Ifng* were filtered out in their respective graphs. **f**, Pseudotime assignment for WT and *Rptor*^{Il17aCre} cells, coloured by genotype; each dot represents one cell. **g**, *Cd27* and *Tcf7* gene expression across pseudotime, coloured by genotype. **h**, tSNE visualization of *Cd27* and *Tcf7* expression. **i**, Flow cytometry analysis of T-bet expression in freshly isolated CD27$^+$ and CD27$^-$ cells from dLN of Il17aCre (R26R^{eYFP}) mice at day nine post MOG-immunization. **j**, Fold change in the percentage of the IL-17$^-$ IFN-γ^+ cells amongst CD27$^+$ or CD27$^-$ YFP$^+$ cells stimulated with MOG plus IL-12 as compared with freshly isolated cells ($n = 12$). **k**, CD27$^+$ YFP$^+$ cells from MOG-immunized WT and *Rptor*^{Il17aCre} mice were sorted and transferred into CD45.1$^+$ hosts. The following day, CD45.1$^+$ host mice were immunized with MOG; four days later, YFP$^+$ cells were analysed by flow cytometry for surface CD27 expression (left; a summary plot is at the right: $n = 6$, WT; $n = 5$, *Rptor*^{Il17aCre}). **l**, CD27$^+$ YFP$^+$ cells from MOG-immunized WT and *Rptor*^{Il17aCre} mice were stimulated with MOG plus IL-12 for four days, and then CD27 expression was analysed. **m**, T_H17 cells are functionally and metabolically heterogeneous, and are composed of a subset with stemness features but lower anabolic metabolism, and a reciprocal subset with higher metabolic activity that supports transdifferentiation into T_H1 cells. These two subsets are further distinguished by selective expression of the transcription factors TCF-1 and T-bet, respectively, and discrete levels of CD27 expression. mTORC1 activation drives reprogramming of anabolic metabolism, favouring transcription that is mediated by T-bet rather than TCF-1; consequently, T_H17 transdifferentiation into T_H1-like T_H17 cells occurs. Memory/stem-like T_H17 cells can become reactivated and have the potential to undergo terminal differentiation and acquire T_H1-like phenotypes. Data are means ± s.e.m. and from one experiment (**a**–**h**), or are representative of three (**i**) or five (**j**–**l**) independent experiments. Numbers in plots represent frequencies of cells in gates; numbers within histograms represent mean fluorescence intensities. Mann–Whitney U-test (two-sided) was used in panel **j** to determine statistical significance.

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0802-y

Programmable design of orthogonal protein heterodimers

Zibo Chen[1,2,3], Scott E. Boyken[1,2], Mengxuan Jia[4], Florian Busch[4], David Flores-Solis[5], Matthew J. Bick[1,2], Peilong Lu[1,2], Zachary L. VanAernum[4], Aniruddha Sahasrabuddhe[4], Robert A. Langan[1,2,3], Sherry Bermeo[1,2,3], T. J. Brunette[1,2], Vikram Khipple Mulligan[1,2], Lauren P. Carter[1,2], Frank DiMaio[1,2], Nikolaos G. Sgourakis[5], Vicki H. Wysocki[4] & David Baker[1,2,6]*

Specificity of interactions between two DNA strands, or between protein and DNA, is often achieved by varying bases or side chains coming off the DNA or protein backbone—for example, the bases participating in Watson–Crick pairing in the double helix, or the side chains contacting DNA in TALEN–DNA complexes. By contrast, specificity of protein–protein interactions usually involves backbone shape complementarity[1], which is less modular and hence harder to generalize. Coiled-coil heterodimers are an exception, but the restricted geometry of interactions across the heterodimer interface (primarily at the heptad a and d positions[2]) limits the number of orthogonal pairs that can be created simply by varying side-chain interactions[3,4]. Here we show that protein–protein interaction specificity can be achieved using extensive and modular side-chain hydrogen-bond networks. We used the Crick generating equations[5] to produce millions of four-helix backbones with varying degrees of supercoiling around a central axis, identified those accommodating extensive hydrogen-bond networks, and used Rosetta to connect pairs of helices with short loops and to optimize the remainder of the sequence. Of 97 such designs expressed in *Escherichia coli*, 65 formed constitutive heterodimers, and the crystal structures of four designs were in close agreement with the computational models and confirmed the designed hydrogen-bond networks. In cells, six heterodimers were fully orthogonal, and in vitro—following mixing of 32 chains from 16 heterodimer designs, denaturation in 5 M guanidine hydrochloride and reannealing—almost all of the interactions observed by native mass spectrometry were between the designed cognate pairs. The ability to design orthogonal protein heterodimers should enable sophisticated protein-based control logic for synthetic biology, and illustrates that nature has not fully explored the possibilities for programmable biomolecular interaction modalities.

Orthogonal sets of protein–protein and protein–peptide interactions have important roles in biological systems[6]. It has proven difficult to use sequence redesign to create new specificities starting from naturally occurring interacting proteins, such as toxin–antidote pairs[7] (promiscuous binding has usually resulted[8]); the natural specificity results at

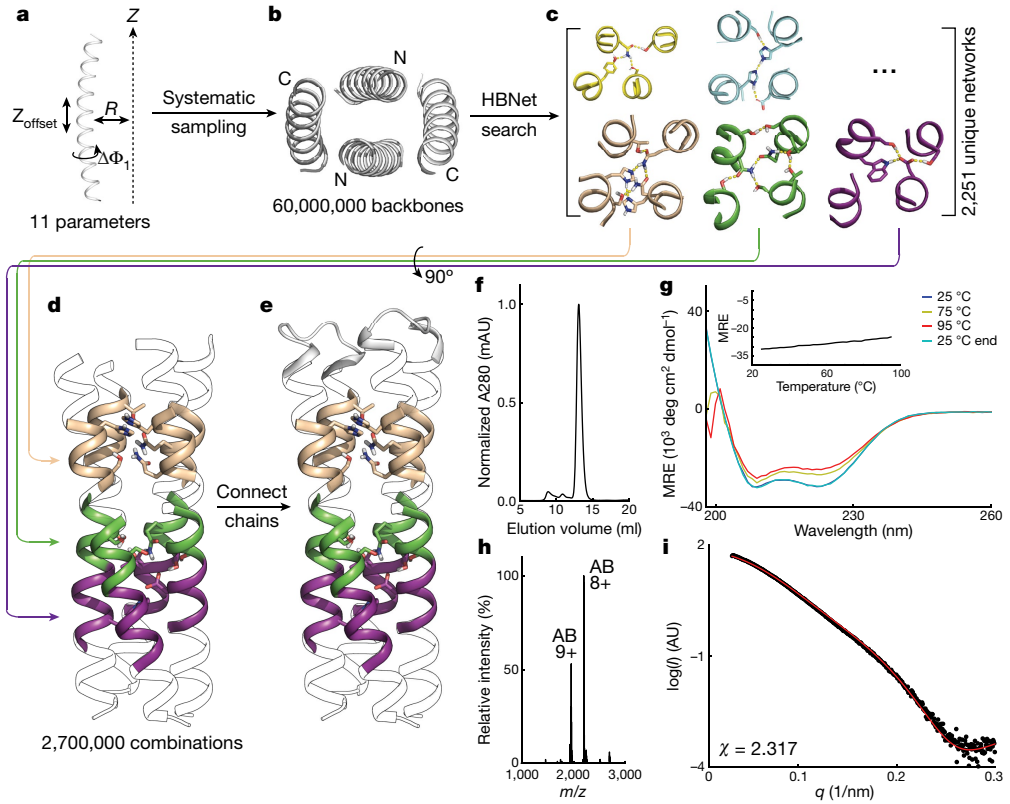


Fig. 1 | Modular heterodimer design. a, Individual helix generation: the helical phase ($\Delta\phi_1$), supercoil radius (R) and offset along the Z-axis (Z_{offset}) were exhaustively sampled; a total of 11 free parameters, because there is no Z_{offset} for the first helix. **b**, Top-down view of a representative four-helix backbone. **c**, Representative hydrogen-bond networks identified using HBNet. **d**, Matches of multiple HBNet-containing heptads to a single full-length backbone. **e**, Addition of loops to connect the four helices into two helix hairpins. **f**–**i**, SEC trace (**f**), circular dichroism spectra and (inset) temperature melt (**g**), native MS spectrum (**h**) and SAXS (**i**; black, experimental SAXS data; red, spectra computed from the designed backbones; q, scattering vector) profiles of the design DHD37_ABXB. Experiments were performed once. A280, absorbance at 280 nm; mAU, milli-absorbance units; MRE, molar residue ellipticity.

[1]Department of Biochemistry, University of Washington, Seattle, WA, USA. [2]Institute for Protein Design, University of Washington, Seattle, WA, USA. [3]Graduate Program in Biological Physics, Structure, and Design, University of Washington, Seattle, WA, USA. [4]Department of Chemistry and Biochemistry, The Ohio State University, Columbus, OH, USA. [5]Department of Chemistry and Biochemistry, University of California Santa Cruz, Santa Cruz, CA, USA. [6]Howard Hughes Medical Institute, University of Washington, Seattle, WA, USA. *e-mail: dabaker@uw.edu

© 2019 Springer Nature Limited. All rights reserved.

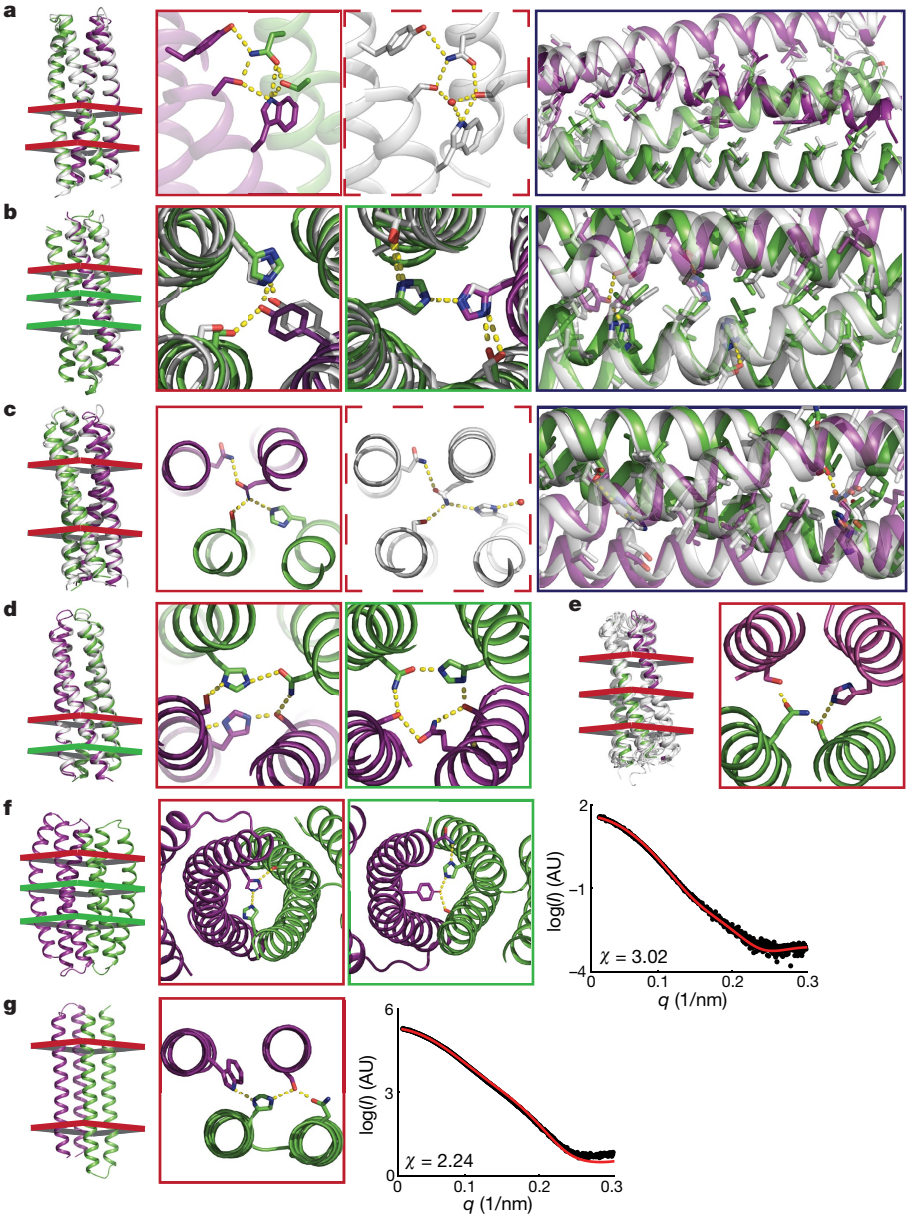


Fig. 2 | Structural characterization of designed heterodimers. **a**–**e**, Crystal and NMR structures (white) superimposed on design models with monomers coloured green and purple; coloured cross-sections of backbones (left) indicate locations of designed hydrogen-bond networks (middle panels). Solid and dashed red boxes compare networks in design model and crystal structure, respectively. Green and black boxes denote additional hydrogen-bond network and hydrophobic packing layers, respectively. **a**, DHD_131, 2.4 Å resolution with 1.0 Å Cα r.m.s.d. **b**, DHD37_1:234, 3.3 Å resolution with 1.4 Å r.m.s.d. **c**, DHD_127, 1.8 Å resolution with 1.7 Å r.m.s.d. **d**, DHD_15, 3.4 Å resolution with 0.9 Å r.m.s.d.; hydrogen-bond networks were not well-resolved. **e**, NMR ensemble (white) of DHD13_XAAA superimposed onto the design model; the assigned side-chain–side-chain NMR distance constraints were not sufficient to define hydrogen-bond networks. **f**, **g**, Backbones and designed hydrogen-bond networks of DHD_39 and DHD_120. Experimental SAXS data (black) are similar to spectra computed from the designed backbones (red).

least in part from complementary variation in backbone conformation[9]. Orthogonal sets of 2–4 interacting coiled-coil pairs have been created and experimentally validated[10,11], including the widely used SYNZIPs[12–18], but interaction promiscuity has again hampered the design of larger orthogonal sets.

Guided by the example of the DNA double helix, we hypothesized that large sets of designed heterodimers could be generated by incorporating asymmetric buried hydrogen-bond networks into regularly repeating backbone structures. We generated helical bundle heterodimers in which each monomer is a helix–turn–helix starting from four-helix backbones produced using a generalization of the Crick coiled-coil parameterization[5,19]. For each of the four helices, we exhaustively sampled the helical phase ($\Delta\phi_1$), supercoil radius (R) and offset along the Z-axis (Z_{offset}) (Fig. 1a), restricting the supercoil phases of the helices to 0, 90, 180 and 270°, and the supercoil twist (ω_0) and helical twist (ω_1) to the ideal values for either a two-layer left-handed supercoil ($\omega_0 = -2.85$ and $\omega_1 = 102.85$), or a five-layer untwisted bundle ($\omega_0 = 0$ and $\omega_1 = 100$)[20] (Extended Data Fig. 1a–d). This yielded 27 million untwisted and 60 million left-handed supercoiled backbones for both parallel and antiparallel orientations of opposing helices (Fig. 1b, Extended Data Fig. 1g).

To identify the modular hydrogen-bond network equivalents to DNA base pairs, we used Rosetta HBNet[21] to design buried hydrogen-bond networks in the central repeat units of each backbone, and obtained 2,251 hydrogen-bond networks involving at least four side-chain residues with all heavy-atom donors and acceptors participating in

© 2019 Springer Nature Limited. All rights reserved.

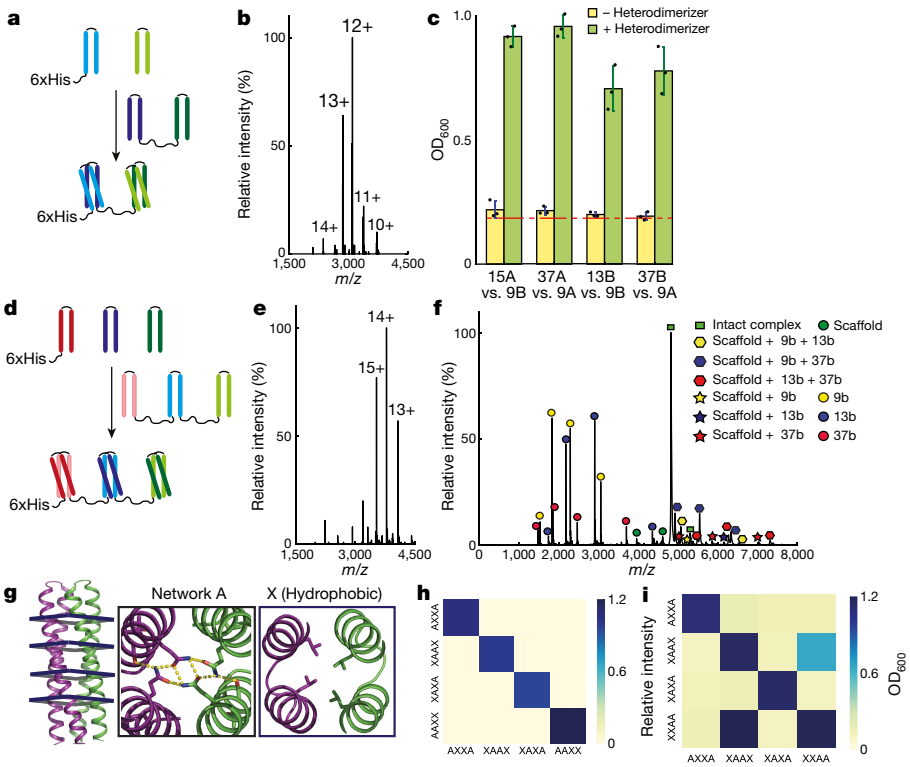


Fig. 3 | New functionality from DHD combinations. a, Induced dimerizer formed from *b* component of DHD13_XAAA (dark blue) fused to *b* component of DHD37_ABXB (dark green) with an intervening flexible linker. The *a* components of the two heterodimers (light blue and light green) are brought into close proximity by the heterodimerizer. **b**, Native MS of purified DHD13_XAAA:DHD37_ABXB heterotrimer complex; no heterodimers or monomers were observed. Molecular mass: designed 37,133 Da; observed 37,132 Da. **c**, Y2H data on four induced dimerization systems. Yellow, without heterodimerizer fusion; green, with heterodimerizer fusion. Red dashed line indicates background growth with unfused activation domain and DBD. Data are mean ± s.d. from three biological repeats. **d**, 9_a (pink), 13_XAAA_a (light blue) and 37_ABXB_a (light green) were covalently linked to form a scaffold, recruiting 9_b (red, hexahistidine tagged), 13_XAAA_b (dark blue) and 37_ABXB_b (dark green). **e**, Native MS of purified scaffold complex; no heterotrimers, heterodimers or monomers were observed. Molecular mass: designed 52,979 Da; observed 52,979 Da. **f**, SID of the 11+ peak in **e**; no cross binding between *b* monomers is detected. **g**, The backbone of 2L4HC2_23 can accommodate hydrogen-bond networks at four heptad positions. **h**, Native MS mixing data of four variants generated by hydrogen-bond-network shuffling; the interactions are orthogonal. **i**, Y2H data of four hydrogen-bond shuffling variants. Two biologically independent experiments were performed for **b**, **e**, **f**, **h**, **i**.

hydrogen bonds, and connecting all four helices (Fig. 1c, Extended Data Fig. 2, Supplementary Table 1). We then identified all of the geometrically compatible placements of these hydrogen-bond networks in each backbone (Fig. 1d), selected backbones that accommodated at least two networks, and connected pairs of helices with short loops (Fig. 1e). Low-energy sequences were identified using RosettaDesign[22] calculations in which the hydrogen-bond networks were held fixed. Designs with fully satisfied hydrogen-bond networks and tight hydrophobic packing were selected for experimental characterization, excluding those with networks with *C*2 symmetry to disfavour homodimerization of monomers (Extended Data Fig. 1e, f). Designed heterodimers (DHDs) are referred to by numbers with monomers labelled *a* or *b*; for example, DHD15_a refers to monomer *a* of design DHD15.

Of the 97 selected designs (Supplementary Table 2), 94 were well-expressed in *E. coli* with both monomers co-purifying by Ni-affinity chromatography (only one monomer contains a hexahistidine tag; Supplementary Table 3). For 85 of these 94, the dominant species observed in size-exclusion chromatography (SEC) had the expected size (Fig. 1f). Thirty-nine of these 85 were exclusive heterodimers at 15 μM by native mass spectrometry (MS)[23,24] (Fig. 1h), 13 were heterodimers with a minor population of heterotetramers, and 13 formed heterodimers with one monomer (but never both) also present as a homodimer arising from unbalanced expression in *E. coli* (Supplementary Tables 4–6). Native MS experiments with serially diluted samples suggested that the DHDs have affinities in the nanomolar range (Supplementary Table 7). Three designs characterized by circular dichroism spectroscopy were found to be all α-helical and stable at 95 °C (Fig. 1g, Extended Data Fig. 3).

We investigated the extent to which the heterodimer set could be expanded by permuting the hydrogen-bond networks in the different helical repeat units, and by permuting the backbone connectivity. Assigning each unique network a letter, DHD37_XBBA indicates a variant in which the second, third and fourth repeat units have hydrogen-bond networks B, B and A, and the first heptad has exclusively hydrophobic residues in the core, whereas DHD103_1:423 indicates a heterodimer in which one monomer consists of the first helix of DHD103 and the other monomer consists of helices 2–4 (Extended Data Fig. 4). Thirteen of fourteen hydrogen-bond-network-permuted variants and nine of ten '3 + 1' backbone-permuted heterodimers (generated from five starting '2 + 2' heterodimers) ran as single peaks on SEC and were constitutive heterodimers by native MS (Fig. 2b, Supplementary Tables 8, 9). Using the hydrogen-bond network permutation approach, we were also able to generate a set of four orthogonal homodimers (Fig. 3g–i, Extended Data Fig. 5, Supplementary Table 10).

Small-angle X-ray scattering (SAXS) spectra collected for 44 designs that were constitutive heterodimers by native MS are consistent with the design models (Figs. 1i, 2f, g, Extended Data Fig. 6, Supplementary Table 11). The nuclear magnetic resonance (NMR) structure of DHD13_XAAA closely matched the design model: the root mean square deviation (r.m.s.d.) over all main-chain α-carbon (Cα) atoms was 2 Å between the designed structure and the lowest-energy NMR model (Fig. 2e). The X-ray crystal structures of

© 2019 Springer Nature Limited. All rights reserved.

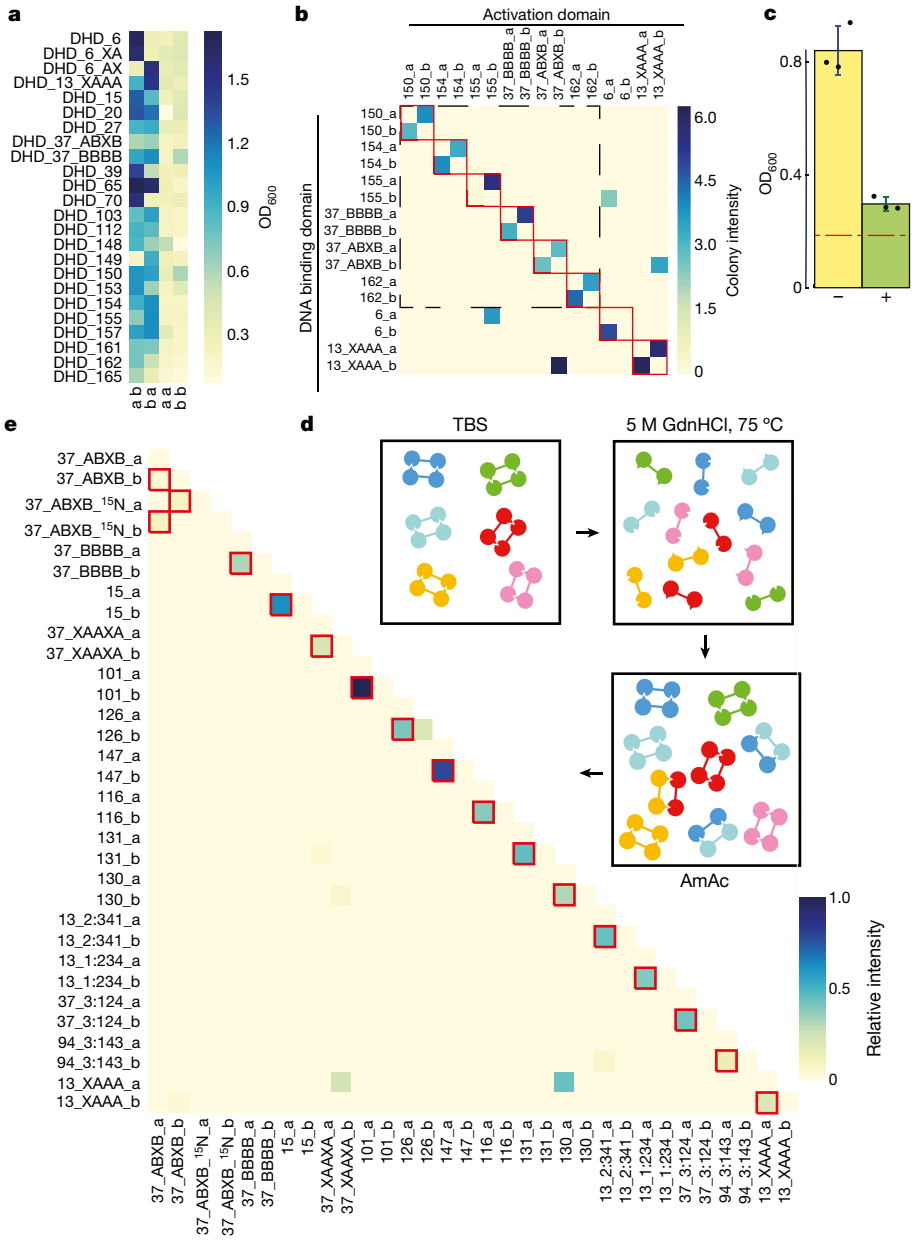


Fig. 4 | All-against-all orthogonality assessment. a, Y2H for 21 heterodimers shows heterodimer formation with little homodimer formation. First letter at bottom indicates monomer fused to activation domain; second letter indicates monomer fused to DBD. **b**, Y2H all-by-all testing of eight pairs of heterodimers; colours indicate growth. Red boxes indicate designed cognate heterodimer pairs, dashed black box indicates a set of six orthogonal heterodimers. **c**, Off-target binding of DHD15_a and DHD13_XAAA_b, in the absence (yellow) or presence (green) of DHD15_b and DHD13_XAAA_a. Data are mean ± s.d. Red dashed line indicates background growth with unfused activation domain and DBD. **d**, **e**, All-against-all orthogonality of 16 pairs of heterodimers assessed by native MS mixing assay. Red boxes indicate designed cognate pairs. Exchange of unlabelled and partially ^{15}N-labelled DHD37_ABXB results in a distribution of overlapping species with low individual signal intensities. Two (**a**, **b**) or three (**c**) biologically independent or three (**e**) technically independent experiments were performed. AmAc, ammonium acetate.

DHD131, DHD37_1:234, DHD127 and DHD15 have backbone Cα atom r.m.s.d. values to the design models ranging from 0.95 to 1.7 Å. The extensive five-residue buried hydrogen-bond network of DHD131 (involving two serines, an asparagine, a tyrosine and a tryptophan) is nearly identical in the crystal structure, with an additional bridging water molecule (Fig. 2a). The two designed hydrogen-bond networks in DHD37_1:234, which contain buried histidine and tyrosine aromatic side chains that sterically disfavour homodimers, are in close agreement with the crystal structure (Fig. 2b). In DHD127, the histidines in the two hydrogen-bond networks adopt a rotamer that differs from the design model (Fig. 2c), making a hydrogen bond with a water molecule. A crystal structure of DHD15 at pH 7.0 is similar to the design model (Fig. 2d), whereas a structure at pH 6.5 is of a domain-swapped, heterotetramer conformation; native MS at pH 6.5 suggests that the designed heterodimer—rather than the heterotetramer—is dominant in solution (Extended Data Fig. 7a–c).

We built three induced dimerization systems by fusing one monomer from each of two different heterodimers via a flexible linker, and testing whether the remaining two monomers from each pair could be brought together by the fusion (Fig. 3a). In each case, the three components copurified by Ni-nitrilotriacetic acid (Ni-NTA) chromatography (one monomer has a hexahistidine tag), and native MS showed they formed constitutive heterotrimers (Fig. 3b); no partial complexes (heterodimers) were observed. Surface-induced dissociation (SID) MS[25,26] resulted in binary complexes consisting of the heterodimerizer bound to either one of the monomers (Supplementary Tables 12, 13),

© 2019 Springer Nature Limited. All rights reserved.

indicating that the interaction between monomers is mediated by the dimerizer fusion. In yeast two-hybrid (Y2H) assays with monomers from two different heterodimers fused to the DNA binding domain (DBD) and transcriptional activation domain, expression of the heterodimerizer fusion as a separate polypeptide chain increased signal substantially over background (Fig. 3c).

We constructed synthetic scaffolding systems[27] by covalently linking the *a* subunits of three DHDs via flexible linkers (Fig. 3d), and co-expressing this scaffold and the three separate *b* subunits, one with a hexahistidine tag, in *E. coli*. Native MS of the purified sample revealed a heterotetramer of all four proteins (Fig. 3e), with SID producing subcomplexes consisting of the scaffold with one or two of the three *b* subunits bound (Fig. 3f); no association between *b* subunits without the scaffold was detected. The scaffold plus monomer assembly is stable at 95 °C and has a guanidine denaturation midpoint of 4 M (Extended Data Fig. 7h, i).

By generating interfaces with many polar groups that are energetically costly to bury without geometrically matched hydrogen-bonding interactions, our design protocol implicitly disfavours non-cognate interactions (explicit negative design to disfavour non-cognate interactions is computationally intractable, given the very large number of possible off-target binding modes). We investigated the interaction specificity of the DHDs in cells using Y2H experiments. For 24 designs, strong interactions were observed by Y2H with the two partners fused to the DBD and activation domain, but not when either partner was fused to both domains; the designed heterodimers, but not the homodimers, form in cells (Fig. 4a). The 24 monomers in 12 of these designs were crossed in an all-by-all Y2H experiment; interactions were observed for all cognate pairs, and 27 of the 552 possible non-cognate interactions (Extended Data Fig. 8). Orthogonality was higher for an eight-DHD subset: of 240 possible non-cognate interactions, only four were observed (Fig. 4b; the interacting polar residues are depicted schematically in Extended Data Fig. 9). Coexpression of unfused monomers eliminated off-target interactions (Fig. 4c); the cognate interactions are evidently stronger than the non-cognate interactions.

To probe all-by-all interactions specificity of the designed monomers when allowed to associate freely in a single pot, purified DHDs were mixed, denatured in 5 M guanidine hydrochloride (GdnHCl) at 75 °C (Fig. 4d), and allowed to reanneal by dialysis. An ^{15}N-labelled variant was added as a control for subunit exchange under denaturing conditions; the hybrid labelled–unlabelled complexes expected if full exchange was taking place were observed in all cases (Extended Data Fig. 10). The resulting mixture was analysed by online ion exchange chromatography coupled to high-resolution native MS (Supplementary Table 14). Sixteen designs (15 unique pairs and the ^{15}N-labelled control) with the highest cognate specificity were pooled together. The native MS results on the 32-chain mixture are notable (Fig. 4e). All 16 designed pairs were recovered, and of the 512 non-cognate binary complexes possible, only 6 were observed. No non-cognate trimers or higher-order oligomers were observed. Several of the orthogonal pairs were generated using hydrogen-bond network shuffling and backbone permutation: the orthogonality of the former is due to positioning of the hydrogen-bond networks as the backbone is fixed, and the orthogonality of the latter, to the connectivity of the chain as the sequence is identical. The differences in orthogonality observed in the native MS mixing and Y2H experiments are likely to stem from the dependence of the former on relative affinity (all monomers are present, and only the lowest-energy complexes form), and the latter, on absolute affinity (only two monomers are present at a time).

Our results demonstrate that the domain of unbounded sets of orthogonal heterodimeric biomolecules constructed from a single repeating backbone is not limited to nucleic acids. Interaction specificity arises from extensive buried hydrogen-bond networks such as the fully connected Tyr-Ser-Trp-Asn-Ser crystallographically confirmed network in Fig. 2a, and heterogeneity in the size of the residues at the designed interface (Extended Data Fig. 8d–i), analogous to the contribution of steric effects to the specificity of Watson–Crick base pairing. The power of native MS to determine interaction specificity in complex mixtures is highlighted by the 32-chain mixing experiment in Fig. 4e; of the large number of possible oligomeric complexes that can be formed from these chains (528 two-chain species, 5,984 three-chain species, and so on), only the 15 designed heterodimers and 6 off-target interactions were observed. The relatively simple encoding of specificity in DNA gave birth to a broad spectrum of new technology, from DNA origami[28] to artificial circuits[29]. Our large set of orthogonal interactions—together with the retention of specificity in the fused monomer systems (the induced dimerizer and scaffold of Fig. 3), and the interaction strength hierarchy illustrated by the cognate interaction competition experiment (Fig. 4c)—open the door to protein-based cellular control circuits with faster response times and better integration with signalling inputs and outputs than current nucleic-acid-based circuitry.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0802-y.

Received: 6 June 2018; Accepted: 8 November 2018;
Published online 19 December 2018.

1. Jones, S. & Thornton, J. M. Principles of protein–protein interactions. *Proc. Natl Acad. Sci. USA* **93**, 13–20 (1996).
2. Harbury, P. B., Zhang, T., Kim, P. S. & Alber, T. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science* **262**, 1401–1407 (1993).
3. Diss, M. L. & Kennan, A. J. Orthogonal recognition in dimeric coiled coils via buried polar-group modulation. *J. Am. Chem. Soc.* **130**, 1321–1327 (2008).
4. Thomas, F., Boyle, A. L., Burton, A. J. & Woolfson, D. N. A set of de novo designed parallel heterodimeric coiled coils with quantified dissociation constants in the micromolar to sub-nanomolar regime. *J. Am. Chem. Soc.* **135**, 5161–5166 (2013).
5. Crick, F. H. C. The Fourier transform of a coiled-coil. *Acta Cryst.* **6**, 685–689 (1953).
6. Zarrinpar, A., Park, S.-H. & Lim, W. A. Optimization of specificity in a cellular protein interaction network by negative selection. *Nature* **426**, 676–680 (2003).
7. Aakre, C. D. et al. Evolving new protein–protein interaction specificity through promiscuous intermediates. *Cell* **163**, 594–606 (2015).
8. Joachimiak, L. A., Kortemme, T., Stoddard, B. L. & Baker, D. Computational design of a new hydrogen bond network and at least a 300-fold specificity switch at a protein–protein interface. *J. Mol. Biol.* **361**, 195–208 (2006).
9. Skerker, J. M. et al. Rewiring the specificity of two-component signal transduction systems. *Cell* **133**, 1043–1054 (2008).
10. Crooks, R. O., Baxter, D., Panek, A. S., Lubben, A. T. & Mason, J. M. Deriving heterospecific self-assembling protein–protein interactions using a computational interactome screen. *J. Mol. Biol.* **428**, 385–398 (2016).
11. Gradišar, H. & Jerala, R. De novo design of orthogonal peptide pairs forming parallel coiled-coil heterodimers. *J. Pept. Sci.* **17**, 100–106 (2011).
12. Thompson, K. E., Bashor, C. J., Lim, W. A. & Keating, A. E. SYNZIP protein interaction toolbox: *in vitro* and *in vivo* specifications of heterospecific coiled-coil interaction domains. *ACS Synth. Biol.* **1**, 118–129 (2012).
13. Reinke, A. W., Grant, R. A. & Keating, A. E. A synthetic coiled-coil interactome provides heterospecific modules for molecular engineering. *J. Am. Chem. Soc.* **132**, 6025–6031 (2010).
14. Acharya, A., Rishi, V. & Vinson, C. Stability of 100 homo and heterotypic coiled-coil a-a′ pairs for ten amino acids (A, L, I, V, N, K, S, T, E, and R). *Biochemistry* **45**, 11324–11332 (2006).
15. Grigoryan, G. & Keating, A. E. Structure-based prediction of bZIP partnering specificity. *J. Mol. Biol.* **355**, 1125–1142 (2006).
16. Gonzalez, L. Jr, Woolfson, D. N. & Alber, T. Buried polar residues and structural specificity in the GCN4 leucine zipper. *Nat. Struct. Biol.* **3**, 1011–1018 (1996).
17. Lumb, K. J. & Kim, P. S. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry* **34**, 8642–8648 (1995).
18. Tatko, C. D., Nanda, V., Lear, J. D. & Degrado, W. F. Polar networks control oligomeric assembly in membranes. *J. Am. Chem. Soc.* **128**, 4170–4171 (2006).
19. Grigoryan, G. & Degrado, W. F. Probing designability via a generalized model of helical bundle geometry. *J. Mol. Biol.* **405**, 1079–1100 (2011).
20. Huang, P.-S. et al. High thermodynamic stability of parametrically designed helical bundles. *Science* **346**, 481–485 (2014).
21. Boyken, S. E. et al. De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity. *Science* **352**, 680–687 (2016).
22. Leaver-Fay, A. et al. ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods Enzymol.* **487**, 545–574 (2011).
23. Ruotolo, B. T. & Robinson, C. V. Aspects of native proteins are retained in vacuum. *Curr. Opin. Chem. Biol.* **10**, 402–408 (2006).

© 2019 Springer Nature Limited. All rights reserved.

24. Sahasrabuddhe, A. et al. Confirmation of intersubunit connectivity and topology of designed protein complexes by native MS. *Proc. Natl Acad. Sci. USA* **115**, 1268–1273 (2018).
25. Zhou, M., Huang, C. & Wysocki, V. H. Surface-induced dissociation of ion mobility-separated noncovalent complexes in a quadrupole/time-of-flight mass spectrometer. *Anal. Chem.* **84**, 6016–6023 (2012).
26. Zhou, M. & Wysocki, V. H. Surface induced dissociation: dissecting noncovalent protein complexes in the gas phase. *Acc. Chem. Res.* **47**, 1010–1018 (2014).
27. Anderson, G. P., Shriver-Lake, L. C., Liu, J. L. & Goldman, E. R. Orthogonal synthetic zippers as protein scaffolds. *ACS Omega* **3**, 4810–4815 (2018).
28. Rothemund, P. W. K. Folding DNA to create nanoscale shapes and patterns. *Nature* **440**, 297–302 (2006).
29. Qian, L. & Winfree, E. Scaling up digital circuit computation with DNA strand displacement cascades. *Science* **332**, 1196–1201 (2011).

Acknowledgements We thank Rosetta@Home volunteers for contributing computing resources; A. Kang for protein crystallization support; B. Sankaran for assistance with diffraction data collection; K. Lau and B. Groves for assistance with Y2H assays; S. Rettie for MS support; S. Ovchinnikov for help with TMalign; M. Marty, M. Bern and A. Norris for assistance with native MS; S. Pennington for making media for Y2H assays; the SIBYLS mail-in SAXS program, supported by the DOE BER IDAT grant (DE-AC02-05CH11231) and ALS-ENABLE (P30 GM124169) for SAXS; and A. Keating, G. Rocklin and N. Woodall for feedback on the manuscript. Additional funding and computing resources are listed in the Supplementary Information.

Reviewer information *Nature* thanks G. Grigoryan, C. Robinson and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions Z.C., S.E.B. and D.B. designed the research. Z.C. and D.B. wrote the manuscript. M.J., F.B., Z.L.V., A.S. and V.H.W. performed native MS experiments and analysed data. L.P.C. prepared proteins for NMR experiments. D.F.-S. and N.G.S. performed NMR experiments. Z.C. wrote the heptad stacking code. S.E.B. improved the HBNet method. V.K.M. wrote the parametric backbone generation code. T.J.B. wrote the loop closure code. Z.C. and S.E.B. carried out design calculations, and R.A.L. and S.B. helped. Z.C., M.J.B., P.L. and F.D. solved crystal structures. All authors discussed results and commented on the manuscript.

Competing interests Z.C., S.E.B., R.A.L., S.B. and D.B. are inventors on US provisional patent application no. 62755264 and patent application WO2017173356A1. D.B. and S.E.B. hold equity in Lyell Immunopharma.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0802-y.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0802-y.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to D.B.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Computational design. *Systematic sampling of parametric helical backbones.* We used a generalization of the Crick coiled-coil parameters[5] to independently sample all four helices of the heterodimers supercoiled around the same axis, as previously described[19–21]. The supercoil twist (ω_0) and helical twist (ω_1) were coupled and ideal values were used[20] with ω_0 and ω_1 held constant among the helices. A left-handed supercoil results from $\omega_0 = -2.85$ and $\omega_1 = 102.85$, and a straight bundle with no supercoiling from $\omega_0 = 0$ and $\omega_1 = 100$. The supercoil phases ($\Delta\phi_0$) for the helices were fixed at 0°, 90°, 180° and 270°, respectively. The offset along the Z-axis (Z_{offset}) for the first helix was fixed to 0 as a reference point, with the rest of the helices independently sampling from −1.51 Å to 1.51 Å, with a step size of 1.51 Å. All helices sampled helical phases ($\Delta\phi_1$) independently, from 0° to 90°, with a step size of 10°. Two of the helices with a $\Delta\phi_0$ separation of 180° sampled the radius from Z-axis (R) from 5 Å to 8 Å, while the other two sampled from 7 Å to 10 Å, all with a step size of 1 Å. Each helix is set to have 35 residues to accommodate 5 heptad repeats. After removing redundant sample points from the overlapping regions of radius sampling, the supercoiled helical bundles contained more than 60 million unique backbones, and the straight helical bundles contained more than 27 million unique backbones.

HBNet search. For each parametrically generated backbone, HBNet[21] was used to search the middle heptad for hydrogen-bond networks that connect all four helices, contain at least four side chains contributing hydrogen bonds, have all heavy atom donors and acceptors satisfied, and span the intermolecular interface. Symmetry was not enforced during the HBNet search. For buried interface positions, only non-charged polar amino acids were considered; for residues that were at the boundary between the protein core and surface, all polar amino acids were considered. A subsequent Rosetta design calculation was performed to optimize hydrophobic packing, with atom pair restraints from HBNet being put on the newly identified hydrogen-bond networks. Finally, a minimization step and side-chain repacking step were performed without atom pair restraints on hydrogen-bonding residues to evaluate how well the networks remained intact in the absence of the constraints. Designs with at most five alanines in the middle heptad and no buried unsatisfied polar heavy atoms were selected for downstream design.

Generating combinations of HBNets with heptad stacking. The purpose of this step is to identify five-heptad backbones (full backbones) that can accommodate at least two HBNets. Instead of generating one-heptad backbones and full backbones separately, searching for HBNets in the one-heptad backbones and aligning them to all full backbones, we reasoned that the heptad stacking method would remain the same if we simply searched for HBNets in the middle heptad on all full backbones, extracted the middle heptads, and aligned them to all full backbones. We therefore extracted the middle heptads containing HBNets, generated all variants of chain ordering, and did pairwise alignment of middle heptads to full backbones using TMalign[30]. All alignments with r.m.s.d. less than 0.3 were identified and full backbones that could accommodate at least two middle heptads were selected for final design.

Connecting parametric helical backbones. Helical backbones are connected with short 2–5-residue loops such that the r.m.s.d. of each loop is less than 0.4 to a 9-residue stretch in a native protein. The distance and directionality between helices limit what loops can connect, so our closure extends and shrinks helices by up to three residues. We then superimpose all short loops from the PDB onto the first and last two helical residues. The loops with the lowest stub-r.m.s.d. are minimized using the Rosetta score function onto the helical endpoints to ensure near-perfect closure. Loop quality is assessed by measuring the distance in r.m.s.d. to the closest nine stretch in the PDB. The loop with the lowest r.m.s.d. is returned as the solution. We repeat this procedure to connect all helices and report the solution with the lowest r.m.s.d.

Design calculations. Backbones were regularized using Cartesian space minimization in Rosetta to alleviate any torsional strain introduced by heptad stacking. Two consecutive Rosetta packing rounds were performed with increasing weight on the repulsive energy to optimize hydrophobic packing, while constraining the hydrogen-bond network residues. A FastDesign step was subsequently used within a generic Monte Carlo mover to optimize secondary structure shape complementarity, while allowing at most 8% alanine, three methionines and three phenylalanines in the protein core. The last step of minimization and side-chain repacking to identify the movement of HBNets without atom pair constraints is the same as described in 'HBNet search' above.

Selection criteria and metrics used to evaluate designs. Designs were selected using the following criteria: change in polar surface area upon binding (dSASA_polar) greater than 800 Å; secondary structure shape complementarity (ss_sc) score greater than 0.65; holes score around HBNets less than −1.4; no buried unsatisfied heavy atoms; at least one buried bulky polar side chain per monomer. Selected designs were then visually inspected for good packing of hydrophobic side chains, especially the interdigitation of isoleucine, leucine and valine. Surface tyrosines were added at non-interfering positions to aid protein concentration measurement by recording optical density at 280 nm (OD_{280}). Surface charge residues for a few of the designs were redesigned to shift the theoretical isoelectric point away from buffer pH.

Calculations of r.m.s.d. Crystal structures and the corresponding design models were superimposed with TMalign using all heavy atoms. From this alignment, r.m.s.d. was calculated across all α-carbon atoms, and also across heavy atoms of the hydrogen-bond network residues.

Logistic regression. Designs were first scored with various filters in Rosetta with the filter values reported. Experimental results and Rosetta filter values were used as inputs to a logistic regression method[31] to find correlations between computational metrics and experimental observations.

Visualization and Figures. All structural images for figures were generated using PyMOL[32].

Buffer and medium recipes. TBM-5052: 1.2% (wt/vol) tryptone, 2.4% (wt/vol) yeast extract, 0.5% (wt/vol) glycerol, 0.05% (wt/vol) D-glucose, 0.2% (wt/vol) D-lactose, 25 mM Na_2HPO_4, 25 mM KH_2PO_4, 50 mM NH_4Cl, 5 mM Na_2SO_4, 2 mM $MgSO_4$, 10 μM $FeCl_3$, 4 μM $CaCl_2$, 2 μM $MnCl_2$, 2 μM $ZnSO_4$, 400 nM $CoCl_2$, 400 nM $NiCl_2$, 400 nM $CuCl_2$, 400 nM Na_2MoO_4, 400 nM Na_2SeO_3, 400 nM H_3BO_3. Lysis buffer: 20 mM Tris, 300 mM NaCl, 20 mM imidazole, pH 8.0 at room temperature. Wash buffer: 20 mM Tris, 300mM NaCl, 30 mM imidazole, pH 8.0 at room temperature. Elution buffer: 20 mM Tris, 300 mM NaCl, 250 mM imidazole, pH 8.0 at room temperature. Buffer W: 100 mM Tris-HCl pH 8.0, 150 mM NaCl and 1 mM EDTA. Buffer E: buffer W containing 2.5 mM D-desthiobiotin. TBS buffer: 20 mM Tris pH 8.0, 100 mM NaCl.

Construction of synthetic genes. For the expression of heterodimers, both monomers were encoded in the same plasmid, separated by a ribosome binding sequence (GAAGGAGATATCATC). Synthetic genes were ordered from Genscript and delivered in pET21-NESG *E. coli* expression vector, inserted between the NdeI and XhoI sites. For the pET21-NESG constructs, a hexahistidine tag and a tobacco etch virus (TEV) protease cleavage site (GSSHHHHHHSSGENLYFQGS) were added in frame at the N terminus of the second monomer. A stop codon was introduced at the 3′ end of the second monomer to stop expression of the C-terminal hexahistidine tag in the vector. For purification with Strep-tactin resin, a streptavidin tag (SAWSHPQFEKGGGSGGGSGGSAWSHPQFEKSGENLYFQGS) coding sequence was cloned in-frame 5′ of the first monomer sequence.

For the co-expression of three and four proteins from the same plasmid (induced dimerization and synthetic scaffold designs), synthetic genes were cloned in the pRSFDuet-1 expression vector. The first (in the case of three proteins) or first two (in the case of four proteins) genes were cloned between NcoI and HindIII sites, with a ribosome binding site separating the two genes in the latter case. The last two genes were cloned between NdeI and XhoI sites, separated by a ribosome binding site. A hexahistidine tag and a TEV protease cleavage site coding sequence were cloned in-frame 5′ of the last gene.

Genes for Y2H studies were cloned into plasmids bearing the GAL4 transcription activation domain (poAD) and the GAL4 DNA-binding domain (poDBD).

Protein expression. Plasmids were transformed into chemically competent *E. coli* expression strains BL21(DE3)Star (Invitrogen) or Lemo21(DE3) (New England Biolabs) for protein expression. Single colonies were picked from agar plates following transformation and growth overnight, and 5-ml starter cultures were grown at 37 °C in Luria-Bertani (LB) medium containing 100 μg/ml carbenicillin (for pET21-NESG vectors) or kanamycin (for pRSFDuet-1 vectors) with shaking at 225 r.p.m. for 18 h at 37 °C. Starter cultures were diluted into 500 ml TBM-5052 containing 100 μg/ml carbenicillin or kanamycin, and incubated with shaking at 225 r.p.m. for 24 h at 37 °C.

For expression of $^{13}C^{15}N$- or ^{15}N-labelled proteins, the plasmids were transformed into the Lemo21(DE3) *E. coli* expression strain and plated on M9/glucose plates containing 50 μg/ml carbenicillin. For the starter culture, a single colony was used for inoculation of 50 ml LB medium with 50 μg/ml carbenicillin in a 250-ml baffled flask, and incubated with shaking at 225 r.p.m. for 18 h at 37 °C. Starter culture (10 ml) was then transferred to a 2-l baffled flask containing 500 ml Terrific Broth (Difco), with 25 mM Na_2HPO_4, 25 mM KH_2PO_4, 50 mM NH_4Cl, 5 mM Na_2SO_4, and 100 μg/ml carbenicillin. The culture was grown at 37 °C to an OD_{600} of approximately 1.0, then centrifuged at 5,000 relative centrifugal force (r.c.f.) for 15 min to pellet the cells. The Terrific Broth medium was removed, and the cells were washed briefly with 30 ml of phosphate buffered saline (PBS). The cells were then transferred to a fresh 2-l baffled flask containing 500 ml labelled medium (25 mM Na_2HPO_4, 25 mM KH_2PO_4, 50 mM $^{15}NH_4Cl$, 5 mM Na_2SO_4, 0.2% (w/v) ^{13}C glucose), and 100 μg/ml carbenicillin. The cells were allowed to grow at 37 °C for 2 h, before isopropyl β-D-1-thiogalactopyranoside (IPTG; Carbosynth) was added to 1 mM and the temperature was reduced to 18 °C. The labelled glucose and NH_4Cl were obtained from Cambridge Isotopes.

© 2019 Springer Nature Limited. All rights reserved.

Affinity purification. Cells were collected by centrifugation for 15 min at 5,000 r.c.f. at 4 °C and resuspended in 20 ml lysis buffer. Lysozyme, DNase, and EDTA-free cocktail protease inhibitor (Roche) were added to the resuspended cell pellet before sonication at 70% power for 5 min. For immobilized metal affinity chromatography, lysates were clarified by centrifugation at 4 °C and 18,000 r.p.m. for at least 30 min and applied to Ni-NTA (Qiagen) columns pre-equilibrated with lysis buffer. The column was washed twice with five column volumes (CV) of wash buffer, followed by 5 CV of elution buffer. For Strep tag purification, elution fractions from immobilized metal affinity chromatography were applied to Strep-Tactin Superflow resin (IBA) pre-equilibrated in buffer W. The column was washed with 5 CV Buffer W, before applying 3 CV buffer E to elute proteins off the column. The mass and purity of eluted proteins were confirmed using electrospray ionization mass spectrometry on a Thermo Scientific TSQ Quantum Access mass spectrometer.

Size-exclusion chromatography. N-terminal hexahistidine tags and streptavidin tags were cleaved with TEV protease overnight at room temperature, at a ratio of 1 mg TEV to 100 mg protein. Prior to addition of TEV, buffer was exchanged into lysis buffer. After TEV cleavage, the sample was passed over an additional Ni-NTA column and washed with 1.5 CV of lysis buffer, and flowthrough was collected and further purified by SEC using a Superdex 75 10/300 increase column (GE Healthcare) in TBS buffer.

Circular dichroism measurements. Circular dichroism (CD) wavelength scans (260–195 nm) and temperature melts (25–95 °C) were performed using an AVIV model 420 CD spectrometer. Temperature melts were carried out at a heating rate of 4 °C/min and monitored by the change in ellipticity at 222 nm; protein samples were diluted to 0.25 mg/ml in PBS pH 7.4 in a 0.1-cm cuvette. GdmCl titrations were performed on the same spectrometer with automated titration apparatus in PBS pH 7.4 at 25 °C, with a protein concentration of 0.025 mg/ml in a 1-cm cuvette with stir bar. Each titration consisted of at least 40 evenly distributed GdmCl concentration points with 1-min mixing time for each step. Titrant solution consisted of the same concentration of protein in PBS + GdmCl.

Nuclear magnetic resonance. SEC-purified $^{13}C^{15}N$-labelled protein was concentrated to >1 mM and buffer exchanged into 50 mM NaCl, 20 mM sodium phosphate, 10% D_2O, 0.01% NaN_3 at pH 6.3. Sample was loaded into a 5.0-mm Shigemi tube and four NMR experiments were recorded and analysed: 2D transverse relaxation optimized spectroscopy-heteronuclear single quantum coherence, 4D HNCH nuclear Overhauser effect spectroscopy (NOESY), 4D HNCH total correlated spectroscopy and 4D HCCH NOESY. The data were acquired with a non-uniform sampling (NUS) scheme and subsequently reconstructed with the SMILE program in nmrPipe. For the NOESY experiments, the mixing time was 120 ms and for the NUS protocol the data were recorded with 0.3% and 3.0% of sparsity for the HNCH and HCCH experiments, respectively. The final spectra were loaded and analysed on Sparky 3.115.

The spin systems were identified using supervised NMR data analysis and 148 residues were successfully assigned (93.67%). The completeness in terms of protons assigned was 87.4% (Supplementary Table 4). For the structural determination, 3,423 peaks were extracted from the NOESY data. Twenty-four long-range contacts ($|i-j|>4$) were manually assigned with the 4D HNCH NOESY experiment and 29 in 4D HCCH. Owing to the lack of stereospecific assignments (ambiguous data), the NOE contacts were considered as non-stereospecific assignments for the methyl groups of Leu and Val residues. Those contacts were principally located at the beginning, centre and end of both sequences. The assignments, chemical shifts and proton–proton constraints were used for RASREC or AutoNoe structural calculations in ROSETTA 3. Summaries of refinement statistics are provided in Supplementary Table 16.

Crystallization of protein samples. Purified protein samples were concentrated to approximately 20 mg/ml in 25 mM Tris pH 8.0 and 150 mM NaCl. Samples were screened with a 5-position deck Mosquito crystal (ttplabtech) with an active humidity chamber, using the following crystallization screens: JCSG+ (Qiagen), Crystal Screen (Hampton Research), PEG/Ion (Hampton Research), PEGRx HT (Hampton Research), Index (Hampton Research) and Morpheus (Molecular Dimensions). The optimal conditions for crystallization of the different designs were found as follows: OPHD_37_N3C1, 0.15 M potassium bromide and 30% w/v polyethylene glycol monomethyl ether 2000; OPHD_127, 0.12 M ethylene glycols, 0.1 M buffer system 3 pH 8,5, and 50% v/v precipitate mix 1 from the Morpheus screen; OPHD_15, 0.2 M ammonium sulfate, 0.1 M BIS-TRIS pH 6.5, 18% v/v polyethylene glycol 400; OPHD_15, 0.1 M imidazole pH 7.0, and 25% v/v polyethylene glycol monomethyl ether 550; OPHD_131, 0.2 M ammonium acetate, 0.1 M HEPES pH 7.5, 25% w/v polyethylene glycol 3,350. Crystals were obtained after 1–14 days by the hanging drop vapour diffusion method with the drops consisting of a 1:1, 2:1 or 1:2 mixture of protein solution and reservoir solution.

X-ray data collection and structure determination. The crystals of the designed proteins were looped and placed in the corresponding reservoir solution, containing 20% (v/v) glycerol if the reservoir solution did not contain cryoprotectant, and flash-frozen in liquid nitrogen. The X-ray datasets were collected at the Advanced Light Source at Lawrence Berkeley National Laboratory with beamlines 8.2.1 and 8.2.2. Datasets were indexed and scaled using either XDS[33] or HKL2000[34]. Initial models were generated by the molecular-replacement method with the program PHASER[35] within the Phenix software suite[36], using the design models as the initial search models. Efforts were made to reduce model bias through refinement with simulated annealing using Phenix.refine[37], or, if the resolution was sufficient, by using Phenix.autobuild[38] with rebuild-in-place set to false, simulated annealing and prime-and-switch phasing. Iterative rounds of manual building in COOT[39] and refinement in Phenix were used to produce the final models. Owing to the high degree of self-similarity inherit in coiled-coil-like proteins, datasets for the reported structures suffered from a high degree of pseudo-translational non-crystallographic symmetry, as reported by Phenix.Xtriage, which complicated structure refinement and may explain the higher-than-expected R values reported. The r.m.s.d. values of bond lengths, angles and dihedrals from ideal geometries were calculated with Phenix[36]. The overall quality of all final models was assessed using the program MOLPROBITY[40]. Summaries of diffraction data and refinement statistics are provided in Supplementary Table 17.

Small-angle X-ray scattering. Samples were purified by SEC in 25 mM Tris pH 8.0, 150 mM NaCl and 2% glycerol; fractions preceding the void volume of the column were used as blanks for buffer subtraction. Scattering measurements were performed at the SIBYLS 12.3.1 beamline at the Advanced Light Source. The X-ray wavelength (λ) was 1.27 Å, and the sample-to-detector distance was 1.5 m, corresponding to a scattering vector q ($q = 4\pi \sin\theta/\lambda$, where 2θ is the scattering angle) range of 0.01 to 0.3 $Å^{-1}$. A series of exposures, in equal sub-second time slices, were taken of each well: 0.3-s exposures for 10 s resulting in 32 frames per sample. For each sample, data were collected for two different concentrations to test for concentration-dependent effects; 'low' concentration samples ranged from 2 to 3 mg/ml and 'high' concentration samples ranged from 5 to 7 mg/ml. Data were processed using the SAXS FrameSlice online serve and analysed using the ScÅtter software package[41,42]. FoXS[43,44] was used to compare design models to experimental scattering profiles and calculate quality of fit (χ) values.

Yeast two-hybrid assay. For each pair of binders tested, chemically competent cells of yeast strain PJ69-4a (MATa trp1-901 leu2-3,112 ura3-52 his3-200 gal4(deleted) gal80(deleted) LYS2::GAL1-HIS3 GAL2-ADE2 met2::GAL7-lacZ) were transformed with the appropriate pair of plasmids containing DBDs or activation domains, using the LiAc/SS carrier DNA/PEG method[45]. In the case of induced dimerization, the heterodimerizer was cloned downstream of one of the 'monomer proteins', separated by a p2a and nuclear locolization sequence (GSGATNFSLLKQAGDVEENPGPGDKAELIPEPPKKKRKVELGTA). The p2a sequence ensures translational cleavage to make the heterodimerizer a separate protein from the monomer protein. The selection of transformed yeast cells was performed in synthetic dropout (SDO) medium lacking tryptophan and leucine for 48 h with shaking at 1,000 r.p.m. at 30 °C. The resulting culture was diluted 1:100 and grown for 16 h in fresh SDO medium lacking tryptophan and leucine, before being transferred to a 96-well plate and diluted 1:100 into SDO medium containing 100 mM 3-amino-1,2,4-triazole (3-AT), lacking tryptophan, leucine and histidine (5 mM 3-AT in the case of induced dimerization). The culture was incubated with shaking at 1,000 r.p.m. at 30 °C. As it is necessary to bring the DBD and the transcription activation domain into proximity for the growth of yeast cells in medium lacking histidine, binding of two proteins was indicated by the growth of yeast cells[46,47]. The optical density of yeast cells was recorded after 48 h. For Y2H assay on agar plates, the 1:100 diluted overnight culture was transferred onto a Nunc OmniTray (Thermo Fisher) using a 96 Solid Pin Multi-Blot Replicator (V&P Scientific), with the agar lacking tryptophan, leucine and histidine, and containing 100 mM 3-AT. The plates were imaged daily until day 5 to monitor the sizes of colonies. Images were analysed by the ColonyArea[48] package on ImageJ.

Native MS assessment of heterodimer affinity. Samples were buffer-exchanged into 200 mM ammonium acetate using Micro Bio-Spin 6 columns (Bio-Rad). Protein concentrations were determined spectroscopically by a NanoDrop 2000c (Thermo Fisher Scientific). After dilution with 200 mM ammonium acetate, proteins were allowed to equilibrate for 24 h at 4 °C. Mass spectra were subsequently recorded by nanoESI-MS using an Exactive Plus EMR Orbitrap instrument (Thermo Fisher Scientific) modified to incorporate a quadrupole mass filter and allow surface-induced dissociation[25,26,49].

Native MS of individual heterodimers. Sample purity and integrity were first analysed using a self-packed buffer exchange column[50] (P6 polyacrylamide gel, BioRad), coupled online to an Exactive Plus EMR Orbitrap instrument (Thermo Fisher Scientific) modified to incorporate a quadrupole mass filter and allow surface-induced dissociation. For online buffer-exchange, 200 mM ammonium acetate, pH 6.8 (AmAc) was used as a mobile phase. Samples that showed specific dimer formation and a good correlation with the theoretical monomer or dimer masses were selected for mixing experiments.

© 2019 Springer Nature Limited. All rights reserved.

Native MS mixing assay and data analysis. In the mixing experiment, heterodimers were mixed in equimolar ratio of 10 μM. GdnHCl was added to a final concentration of 5 M and the mixture was incubated at 75 °C for 30 min to ensure complete denaturation. To allow relative quantification of exchanged species, a control mixing experiment was performed in which the denaturation and refolding steps were omitted. The mixtures were then dialysed against 150 mM AmAc solution for refolding and subsequent formation of protein–protein interactions. Eight microlitres of sample was injected on a ProPac WCX-10 column and separately, a ProPac WAX-10 column (Thermo Scientific) and separated using a Dionex UltiMate 3000 HPLC (Thermo Scientific) by a salt gradient elution from 20 mM AmAc to 1,000 mM AmAc over a period of 55 min. The eluting proteins were detected online using a modified Exactive Plus EMR Orbitrap mass spectrometer. LC–MS analysis was performed for mixtures in full MS mode (no collision voltage applied) and all-ion fragmentation (MSMS) mode with high energy collision-induced dissociation (HCD) 100 V and surface-induced dissociation (SID) 85 V, respectively[51]. Details of instrument settings are in Supplementary Table 15. Data were deconvoluted using Xcalibur (Thermo Scientific), UniDec[52] and Intact Mass (Protein Metrics[53]). The detailed deconvolution parameters are listed in Supplementary Table 15. The deconvoluted mass lists from Intact Mass were searched against a theoretical mass list of all possible monomers to tetramers combinations. Only one trimeric species was found, which corresponds to the cognate 13_XAAA_b + 13_2:341_a + 13_1:234_a heterotrimer formation, which constitutes the 13_XAAA design with the two helices of its *a* monomer coming from 13_2:341_a and 13_1:234_a. Dimers were identified using the full MS runs and MSMS runs with both subunits being detected at the same retention time. The mass tolerance was set to 2 Da and intensity tolerance was set to 1% of the highest intensity. The relative intensity was calculated using the equation

$$I_r(N_aM_b) = \frac{I_{DN}(N_aM_b)}{2I_N(N_aN_b)} + \frac{I_{DN}(N_aM_b)}{2I_N(M_aM_b)}$$

in which $I_r(N_aN_b)$ is the relative intensity of a dimer N_aN_b identified in the mixing experiment. $I_{DN}(N_aN_b)$ is the intensity of the N_aN_b species in the run involving denaturation and refolding. $I_N(N_aN_b)$ and $I_N(M_aM_b)$ are the intensities of the cognate pairs N_aN_b and M_aM_b in the run that skipped denaturation and refolding. The native MS mixing workflow is shown in Extended Data Fig. 10.

Native MS of higher-order hetero-oligomers. Samples were buffer exchanged into 200 mM ammonium acetate using Micro Bio-Spin 6 columns (Bio-Rad). Twenty per cent (v/v) 200 mM triethylammonium acetate (Sigma) was added for charge reduction. SID was performed on an in-house modified SYNAPT G2 HDMS (Waters) with a SID device incorporated between a truncated trap travelling wave ion guide and the ion mobility cell[25]. The following instrument parameters were used: sampling cone, 20 V; extraction cone, 2 V; source temperature, 20 °C; trap gas flow, 2 ml/min; trap bias, 45 V. The SID settings are listed in Supplementary Table 15.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

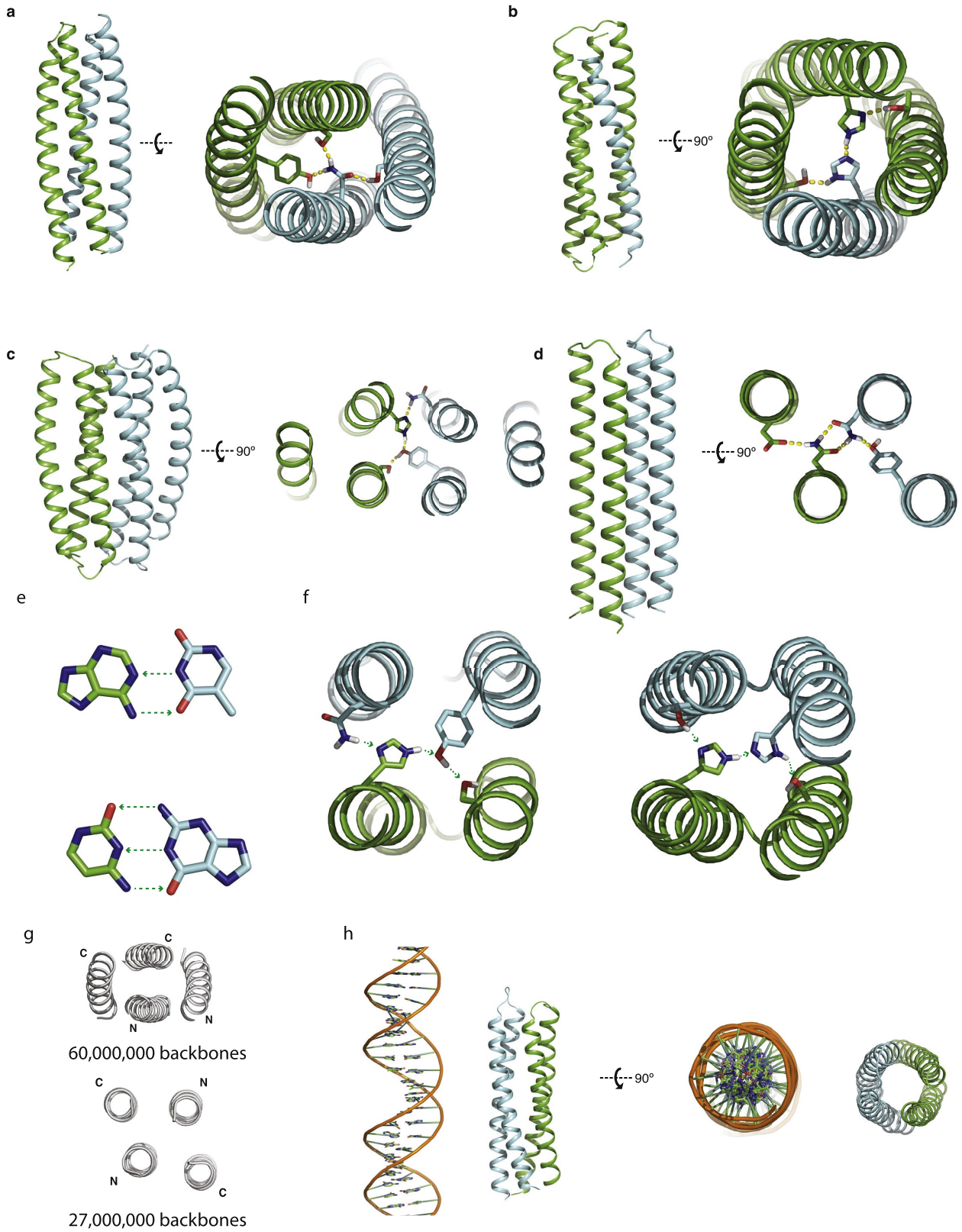
Code availability. All program code is in Rosetta or can be downloaded from the Github repository at https://github.com/uagaug/DeNovoHeterodimers.

Data availability

Coordinates and structure files have been deposited in the Protein Data Bank with accession codes: 6DMP (DHD13_XAAA), 6DKM (DHD131), 6DLC (DHD37_1:234), 6DLM (DHD127), 6DMA (DHD15 heterodimer) and 6DM9 (DHD15 heterotetramer). The native MS spectra generated and analysed during the current study are available at http://files.ipd.uw.edu/pub/de_novo_heterodimers_2018/180813_native_ms_raw.zip. Raw X-ray diffraction images have been deposited at https://proteindiffraction.org/. All source data are available upon request.

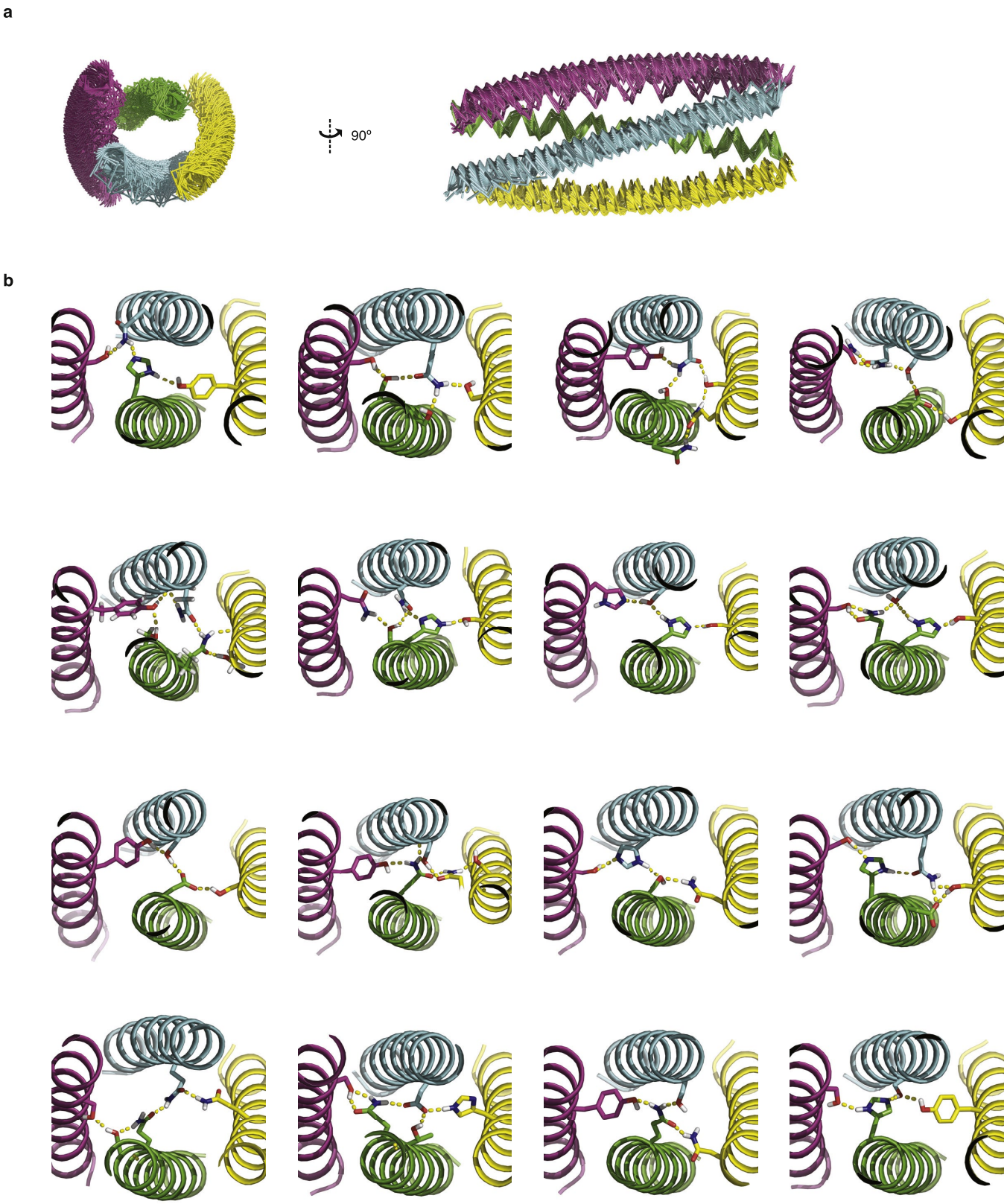
30. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
31. Rocklin, G. J. et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science* **357**, 168–175 (2017).
32. Schrödinger. The PyMOL Molecular Graphics System, Version 1.8. (2015).
33. Kabsch, W. XDS. *Acta Crystallogr. D Biol. Crystallogr.* **66**, 125–132 (2010).
34. Otwinowski, Z. & Minor, W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol.* **276**, 307–326 (1997).
35. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
36. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
37. Afonine, P. V. et al. Joint X-ray and neutron refinement with phenix.refine. *Acta Crystallogr. D* **66**, 1153–1163 (2010).
38. Terwilliger, T. C. et al. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Crystallogr. D* **64**, 61–69 (2008).
39. Emsley, P. & Cowtan, K. Coot: model-building tools for molecular graphics. *Acta Crystallogr. D* **60**, 2126–2132 (2004).
40. Davis, I. W. et al. MolProbity: all-atom contacts and structure validation for proteins and nucleic acids. *Nucleic Acids Res.* **35**, W375–W383 (2007).
41. Dyer, K. N. et al. High-throughput SAXS for the characterization of biomolecules in solution: a practical approach. *Methods Mol. Biol.* **1091**, 245–258 (2014).
42. Rambo, R. P. & Tainer, J. A. Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers* **95**, 559–571 (2011).
43. Schneidman-Duhovny, D., Hammel, M. & Sali, A. FoXS: a web server for rapid computation and fitting of SAXS profiles. *Nucleic Acids Res.* **38**, W540–W544 (2010).
44. Schneidman-Duhovny, D., Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **105**, 962–974 (2013).
45. Schiestl, R. H. & Gietz, R. D. High efficiency transformation of intact yeast cells using single stranded nucleic acids as a carrier. *Curr. Genet.* **16**, 339–346 (1989).
46. Chien, C. T., Bartel, P. L., Sternglanz, R. & Fields, S. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl Acad. Sci. USA* **88**, 9578–9582 (1991).
47. Bartel, P. L., Roecklein, J. A., SenGupta, D. & Fields, S. A protein linkage map of *Escherichia coli* bacteriophage T7. *Nat. Genet.* **12**, 72–77 (1996).
48. Guzmán, C., Bagga, M., Kaur, A., Westermarck, J. & Abankwa, D. ColonyArea: an ImageJ plugin to automatically quantify colony formation in clonogenic assays. *PLoS ONE* **9**, e92444 (2014).
49. Dyachenko, A. et al. Tandem native mass-spectrometry on antibody-drug conjugates and submillion Da antibody–antigen protein assemblies on an orbitrap EMR equipped with a high-mass quadrupole mass selector. *Anal. Chem.* **87**, 6095–6102 (2015).
50. Waitt, G. M., Xu, R., Wisely, G. B. & Williams, J. D. Automated in-line gel filtration for native state mass spectrometry. *J. Am. Soc. Mass Spectrom.* **19**, 239–245 (2008).
51. VanAernum, Z. et al. Surface-induced dissociation of noncovalent protein complexes in an extended mass range Orbitrap mass spectrometer. Preprint available at https://doi.org/10.26434/chemrxiv.7415603.v1 (2018)
52. Marty, M. T. et al. Bayesian deconvolution of mass and ion mobility spectra: from binary interactions to polydisperse ensembles. *Anal. Chem.* **87**, 4370–4376 (2015).
53. Bern, M. et al. Parsimonious charge deconvolution for native mass spectrometry. *J. Proteome Res.* **17**, 1216–1226 (2018).
54. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).

© 2019 Springer Nature Limited. All rights reserved.



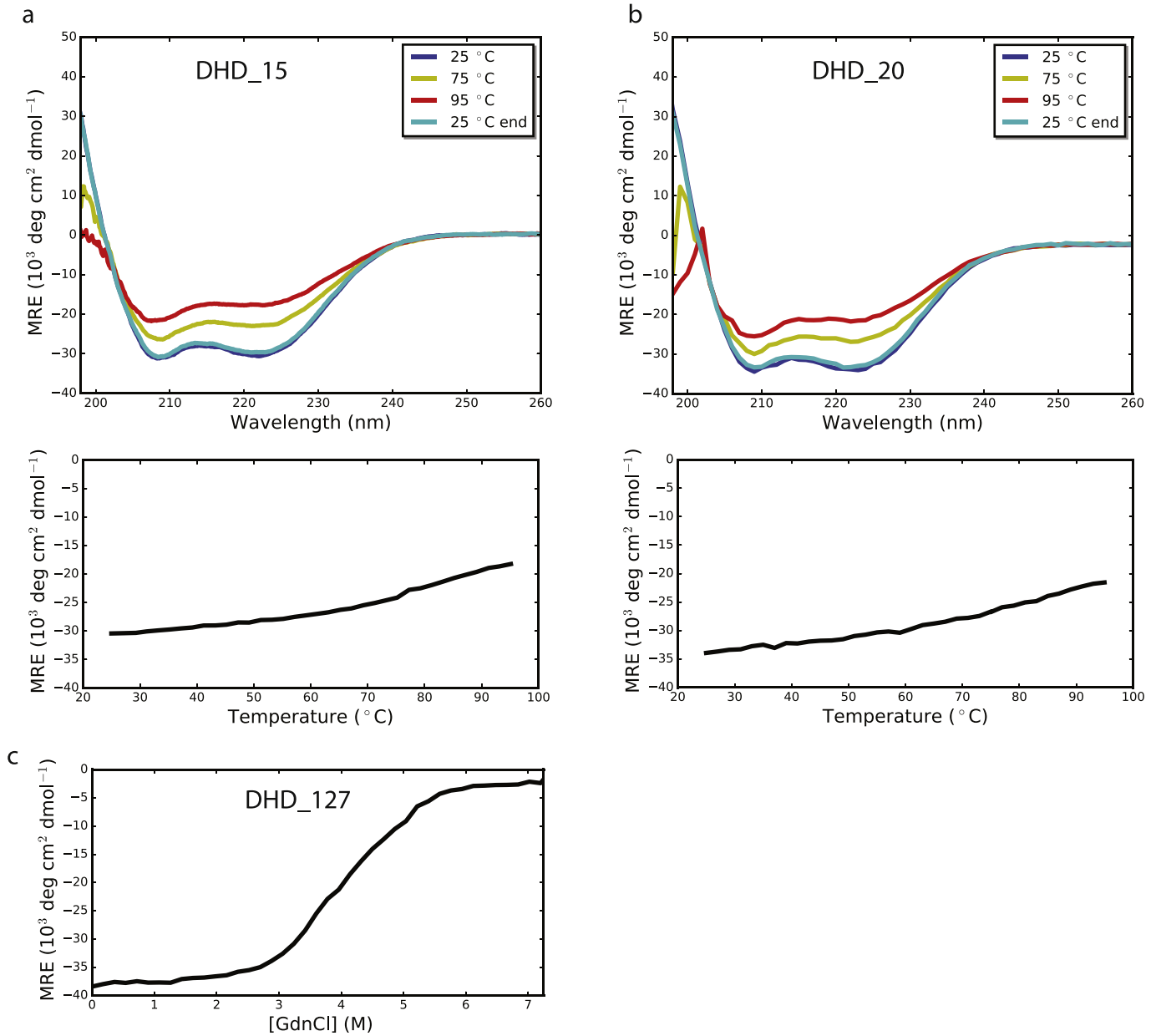
Extended Data Fig. 1 | Overview of different topologies designed. **a**–**d**, Overall topologies on the left and example HBNets on the right. **a**, A left-handed supercoiled backbone, with each monomer being helix hairpins. **b**, A backbone-permuted '3 + 1' design; one monomer is a single helix and the other is a three-helix bundle. **c**, A left-handed supercoiled backbone, with each monomer being a three-helix bundle. **d**, A straight, untwisted backbone, with each monomer being a helix hairpin. **e**, Hydrogen-bond pairing in DNA bases. Top, A–T base pairing. Bottom, C–G base pairing. Green arrows point from hydrogen-bond donors to acceptors. **f**, Two examples of hydrogen-bond pairing in designed protein hydrogen-bond networks. **g**, Top-down view of antiparallel twisted (top) and parallel untwisted (bottom) backbones sampled in this study. **h**, Comparison of a designed protein heterodimer (right) with B-form DNA (left) on the same scale.

© 2019 Springer Nature Limited. All rights reserved.



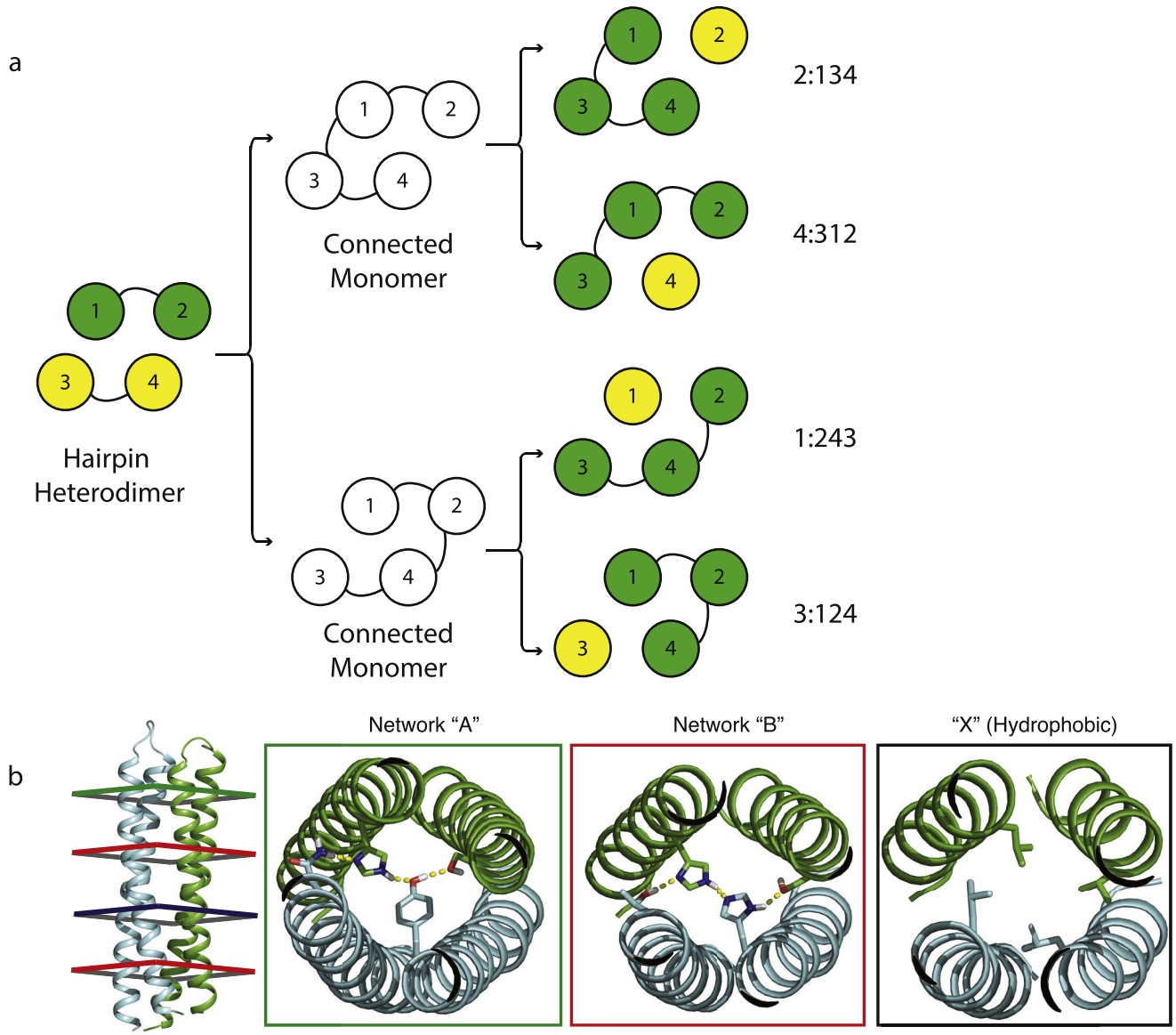
Extended Data Fig. 2 | Example HBNets resulting from the systematic search. a, Overlay of 50 backbones with different Crick parameters for each helix. **b**, Example hydrogen-bond networks from the systematic search, each involving at least four residues and contacting all four helices.

© 2019 Springer Nature Limited. All rights reserved.



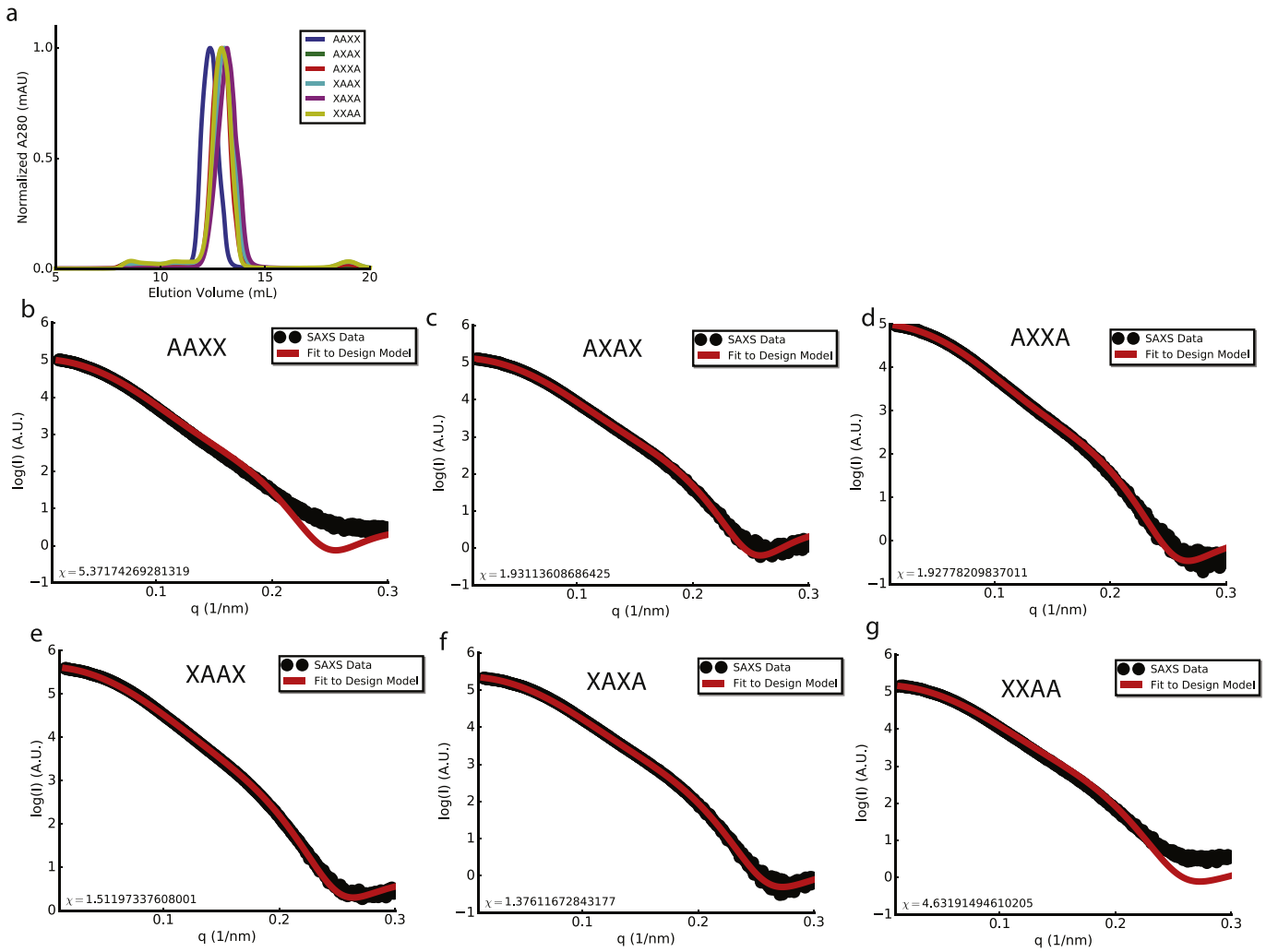
Extended Data Fig. 3 | Thermal and chemical denaturation of DHDs. **a**, **b**, CD spectra for thermal denaturation of DHD_15 and DHD_20, respectively. Top, wavelength scan at 25 °C, 75 °C, 95 °C and final 25 °C. Designs were α-helical and stable up to 95 °C. Bottom, CD temperature melts, monitoring absorption at 222 nm as temperature was increased from 25 °C to 95 °C. **c**, GdnHCl denaturation of DHD_127 measured by CD monitoring absorption at 222 nm. All CD experiments were performed once.

© 2019 Springer Nature Limited. All rights reserved.



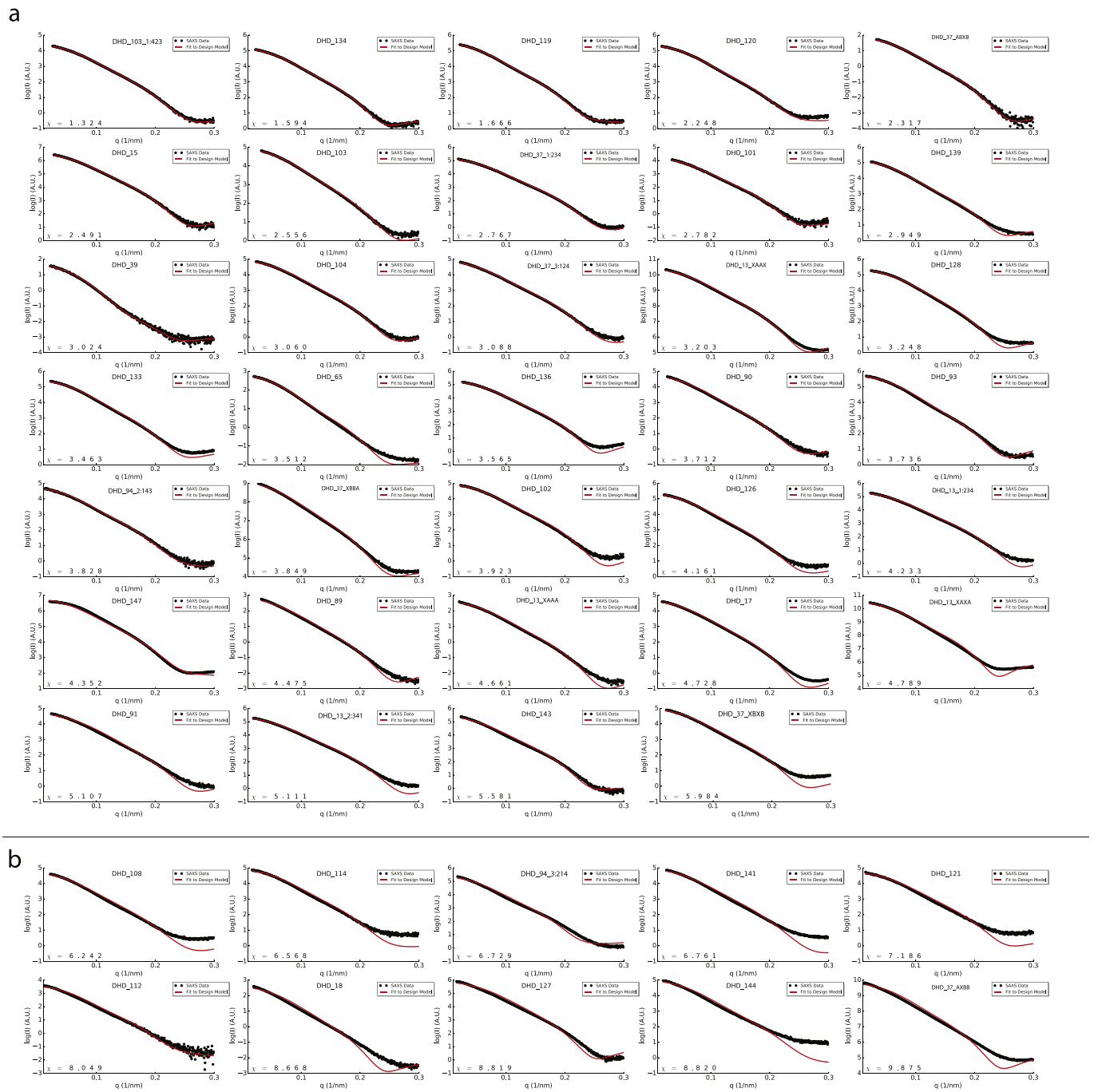
Extended Data Fig. 4 | Backbone and hydrogen-bond network permutations. a, On a 2 + 2 backbone (left), two loops were designed to connect the four helices into a single monomer in two different ways (middle), after which four different cut points were introduced to generate four possible backbone-permuted heterodimers of a single helix and a three helix bundle (3 + 1 heterodimers, right). For example, 2:134 refers to a heterodimer in which the original helix 2 is a single helix, and helices 1, 3 and 4 were connected into a three-helix bundle. **b**, Hydrogen-bond network permutation. Each unique network was assigned a letter (networks 'A' and 'B' in this case), with the hydrophobic packing assigned X. The backbone on the left reads 'ABXB'; its first heptad accommodates network A, its second and fourth heptad accommodate network B, and its third heptad accommodates hydrophobic packing only (X).

© 2019 Springer Nature Limited. All rights reserved.



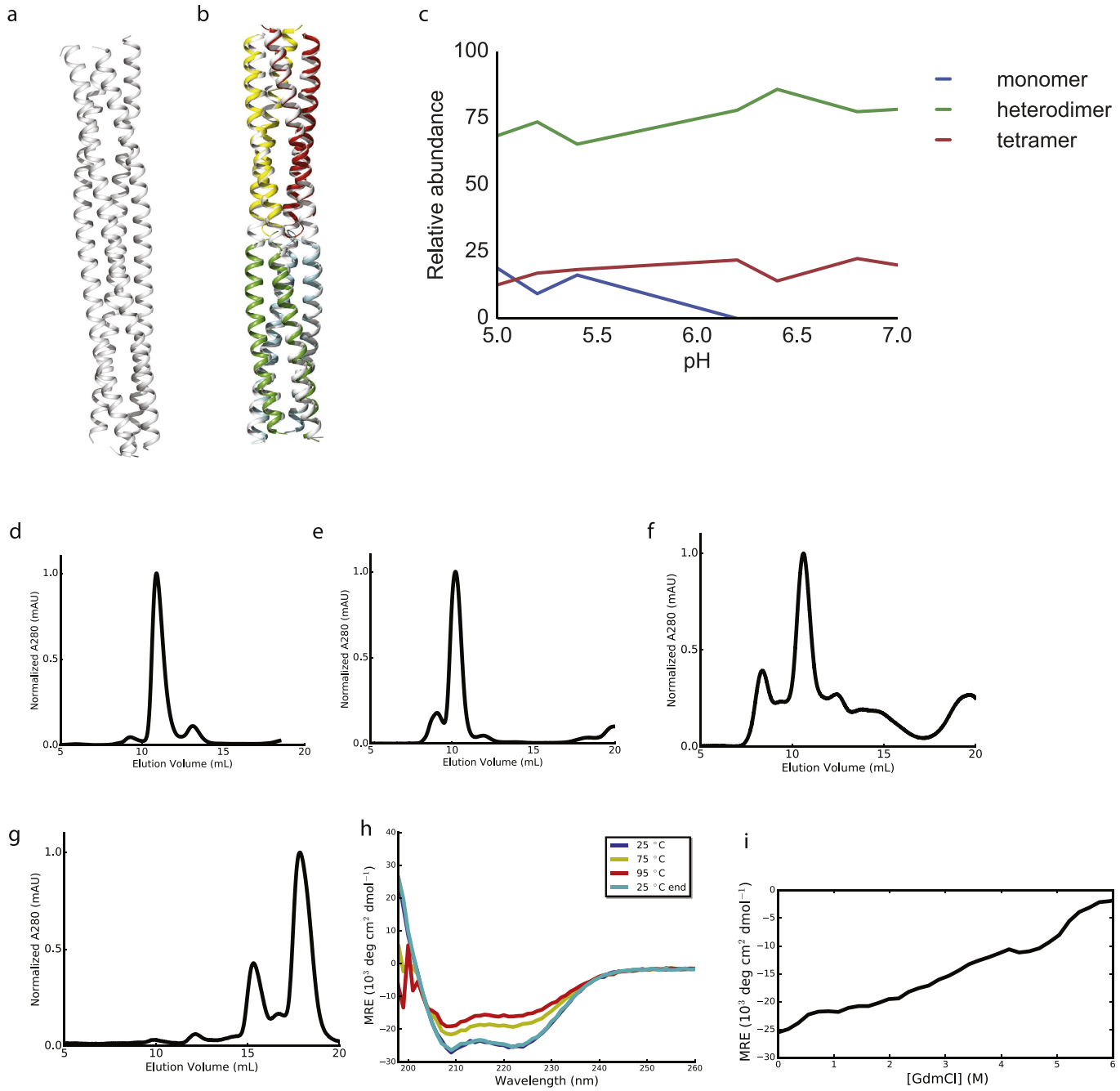
Extended Data Fig. 5 | Biophysical characterization of hydrogen-bond-network-permuted homodimers. a, SEC traces of all six homodimer designs. **b**–**g**, SAXS profiles of hydrogen-bond network-permuted homodimer designs. Black, experimental SAXS data; red, spectra computed from the designed backbones. Two (**a**) or one (**b**–**g**) biologically independent repeats were performed.

© 2019 Springer Nature Limited. All rights reserved.



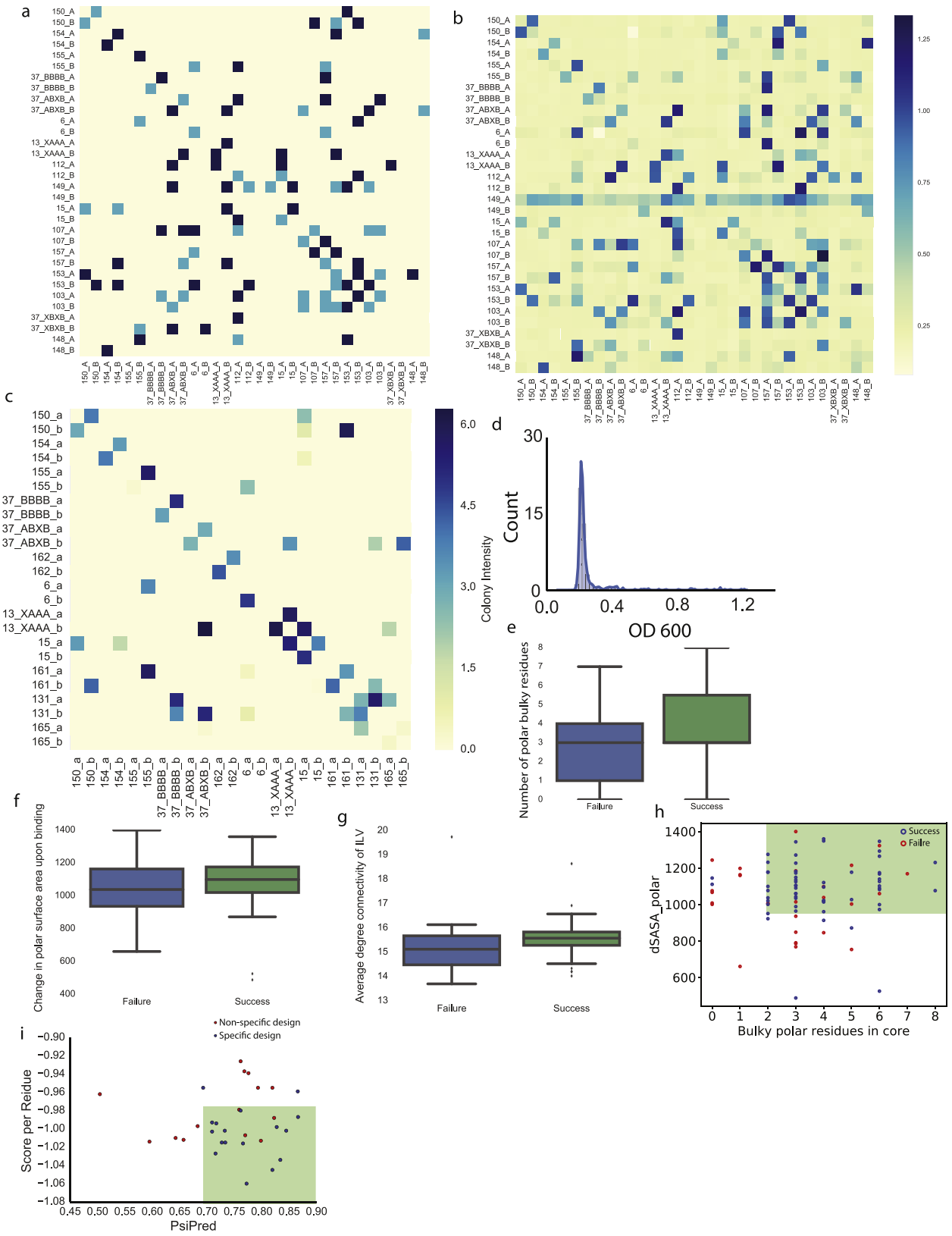
Extended Data Fig. 6 | SAXS profiles of all tested DHDs. Black, experimental SAXS data; red, spectra computed from the designed backbones. **a**, SAXS profiles with χ values smaller than 6. **b**, SAXS profiles with χ values greater than 6. All tested designs showed close agreement to expected radius of gyration (R_g) and maximum distance (d_{max}).

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 7 | Crystal structure of the domain-swapped DHD_15 and biophysical characterization of higher-order oligomers. **a**, Crystal structure of DHD_15 at pH 6.5, with 2.25 Å resolution. **b**, Superposition of design models (in colour) onto both halves of the crystal structure (in white), with backbone r.m.s.d. of 1.83 Å. **c**, Native MS study of DHD_15 at different pH values indicates that heterodimers, rather than heterotetramers, are dominant in solution. **d**–**g**, SEC traces of the induced dimerization DHD_9-13 fusion (**d**), DHD_15-37 fusion (**e**), DHD_13-37 fusion (**f**), and the scaffolding complex in Fig. 3d (**g**; the peak at around 15 ml corresponds to the fully assembled complex, followed by a peak representing an excess of individual components). **h**, CD thermal melt curves for the scaffolding complex in Fig. 3d. Wavelength scan was performed at 25 °C, 75 °C, 95 °C and final 25 °C. Design was α-helical and stable up to 95 °C. **i**, CD chemical denaturation profile of the scaffolding complex in Fig. 3d. Two (**c**–**g**) or one (**h**, **i**) biologically independent repeats were performed.

© 2019 Springer Nature Limited. All rights reserved.

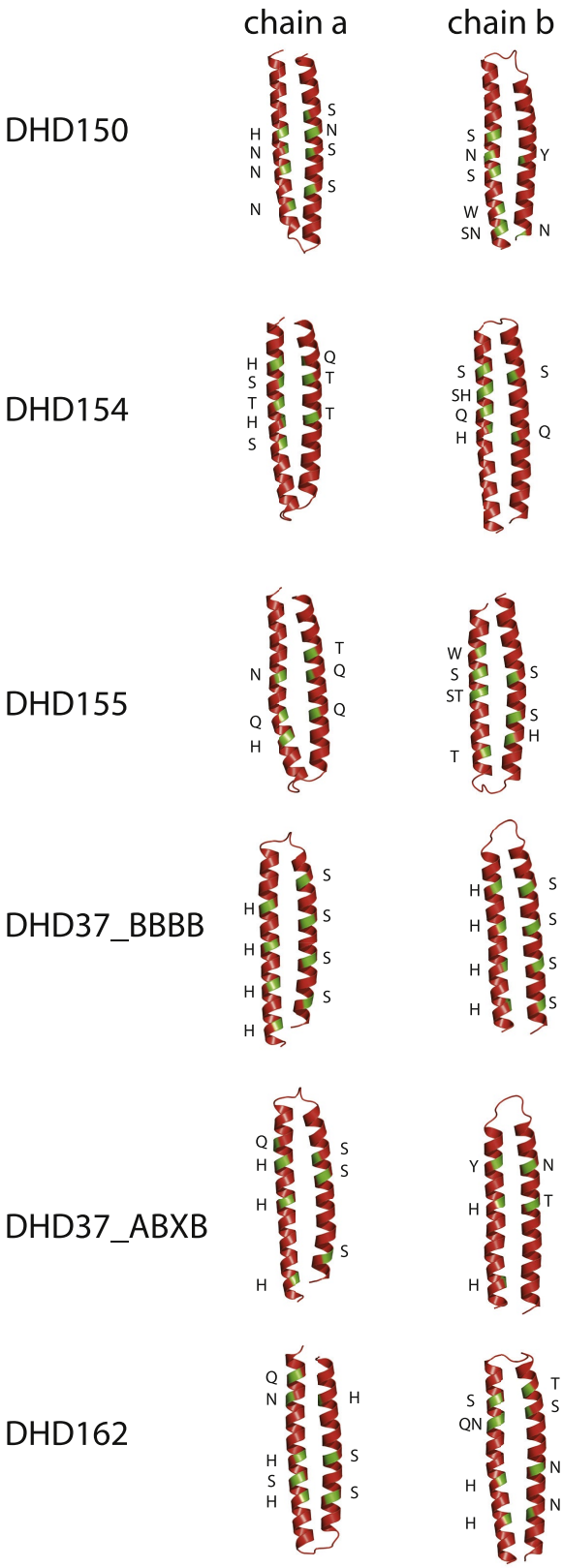


Extended Data Fig. 8 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

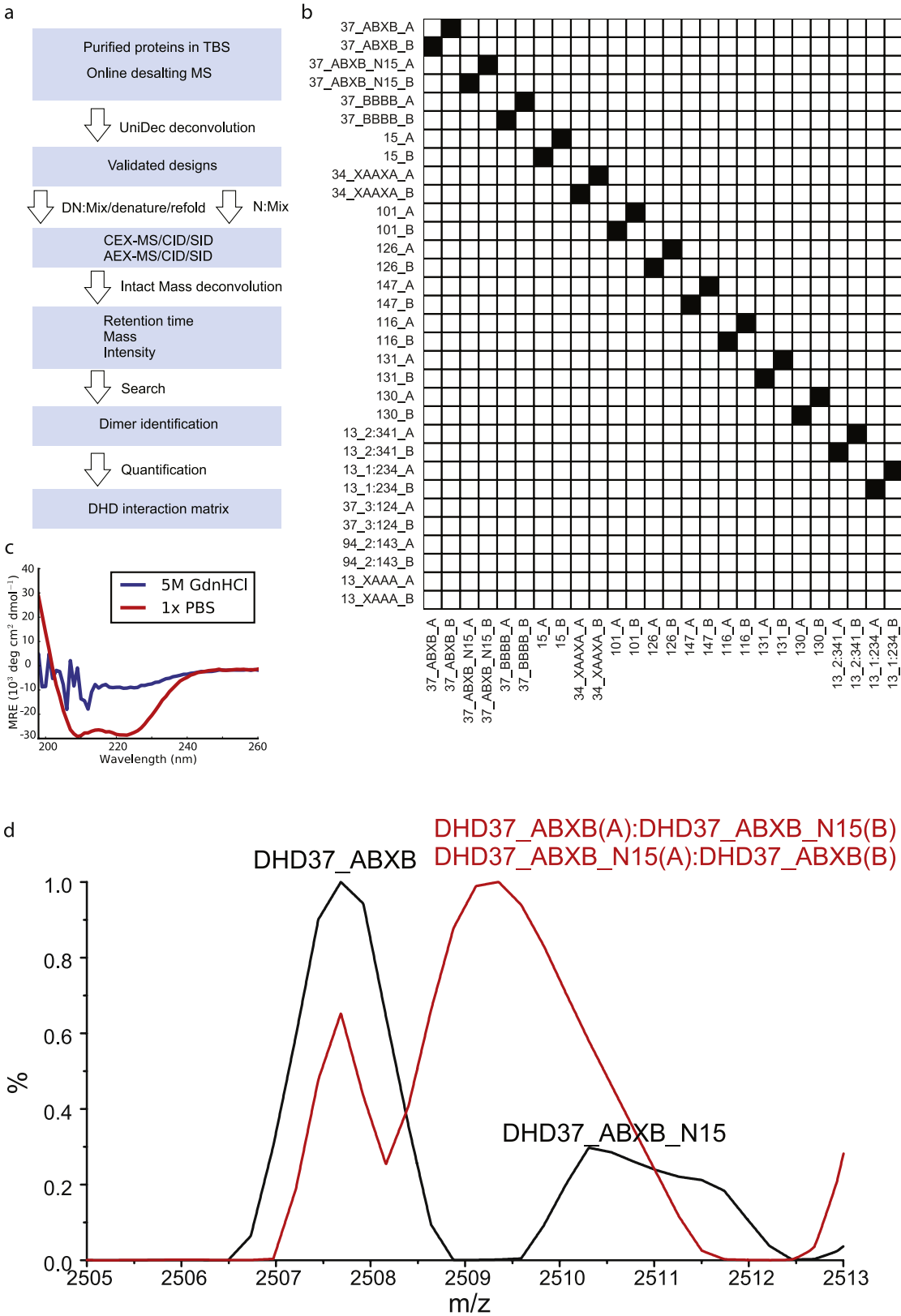
Extended Data Fig. 8 | Y2H all-against-all assay of 16 DHDs. a, Y2H assay with cell growth on agar plates containing 100 mM 3-AT, lacking tryptophan, leucine and histidine. Plates were imaged on day 5. Yellow, no growth on agar plates; light blue, weak growth forming non-circular colonies; dark blue, strong growth. **b**, Y2H result by growing yeast culture in liquid medium containing 100 mM 3-AT, lacking tryptophan, leucine and histidine. OD_{600} values were measured on day 2 to evaluate cell growth. **c**, An additional set of DHDs tested by Y2H showing improved orthogonality. **d**, Distribution of OD_{600} values for non-cognate interactions in **b**. The majority of cells grew to $OD_{600} < 0.4$, indicating weak interactions for non-cognate binding. **e**–**g**, Box plots of various properties for designs that assembled to off-target oligomeric states by native MS (failure) and that assembled into constitutive heterodimers (success). $n = 88$; 25th, 50th and 75th percentiles are shown in the box with the centre being median, extended to 1.5 × interquartile range (IQR) beyond the box. **e**, The number of buried bulky polar residues correlates strongly with design success. **f**, Successful designs tend to have a bigger polar interface surface area. **g**, Designs with better hydrophobic packing (as reported by the Rosetta filter value Average Degree on Ile, Leu and Val residues) tend to have a higher chance of being constitutive heterodimers as assessed by native MS. **h**, Contribution of bulky residues and hydrogen-bond networks to specific dimer formation. dSASA_polar measures interface hydrophilicity and correlates positively with the surface area of hydrogen-bond networks at the interface. Bulky polar residues in core counts the total number of buried bulky residues that participate in hydrogen-bond networks. Constitutive heterodimer formation (blue circles) or off-target oligomer formation (red circles) were determined with native MS. Filter cutoff values of dSASA_polar > 970 Å^2 and more than one polar bulky residue buried in the core includes most of the successful designs and excludes most of the design failures. **i**, On the basis of the Y2H data in **b**, all 32 monomers from the 16 pairs were categorized as being specific (blue, has $\leq$1 non-cognate binding), or non-specific (red, has $>$1 non-cognate binding). With application of secondary structure prediction scores (PsiPred[54]) and Rosetta centroid energy score per residue as filters, designs with higher PsiPred values and lower Rosetta centroid score per residue are more specific (green box). Two independent experiments were performed (**a**–**c**).

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Hydrogen-bond network sequence motifs of the set of six orthogonal pairs in Y2H experiments. Green patches mark the locations of hydrogen-bond network-forming residues on the backbones. Letters along the backbones indicate residue identities.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 10 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Fig. 10 | The workflow of native MS mixing experiments. a, Protein samples were characterized using online desalting coupled to native MS and deconvoluted using UniDec software. Proteins showing expected masses were mixed in equimolar ratio, and the final mix was divided into two parts: in the experimental group (DN), proteins were denatured by 5 M GdnHCl at 75 °C and refoled into 150 mM AmAc; in the control mixing experiment (N), denaturation and refolding steps were omitted. Sample mixtures in each group were further equally divided into three parts that were individually injected on LC–MS with cation exchange and anion exchange, respectively, coupled with CID or SID. LC–MS analysis was performed for mixtures in full MS mode and MSMS mode with HCD and SID, respectively. Data were deconvoluted using Intact Mass. The deconvoluted mass lists from Intact Mass were searched against a theoretical mass list of all possible monomer, dimer, trimer and tetramer combinations. Dimers were identified using the full MS runs and MSMS runs with both subunits being detected at the same retention time. **b**, In the control mixing experiment (N), after mixing all 16 proteins in solution without the denaturation and renaturation steps, no exchange among proteins were observed. **c**, CD data for a mixture of purified DHDs in PBS (red) or 5 M GdnHCl and 75 °C (blue). Protein mixture was fully denatured under the latter conditions. **d**, A mixing experiment of DHD_37_ABXB and ^{15}N-labelled DHD_37_ABXB with (red) or without (black) the denaturation and refolding steps. MS peaks merged after subunit exchange owing to the similarity in the masses of ^{15}N-labelled and unlabelled subunits. Two biologically independent experiments were performed (**b**–**d**).

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0781-z

Trapping biosynthetic acyl-enzyme intermediates with encoded 2,3-diaminopropionic acid

Nicolas Huguenin-Dezot[1,4], Diego A. Alonzo[2,4], Graham W. Heberlig[3], Mohan Mahesh[1], Duy P. Nguyen[1], Mark H. Dornan[3], Christopher N. Boddy[3], T. Martin Schmeing[2]* & Jason W. Chin[1]*

Many enzymes catalyse reactions that proceed through covalent acyl-enzyme (ester or thioester) intermediates[1]. These enzymes include serine hydrolases[2,3] (encoded by one per cent of human genes, and including serine proteases and thioesterases), cysteine proteases (including caspases), and many components of the ubiquitination machinery[4,5]. Their important acyl-enzyme intermediates are unstable, commonly having half-lives of minutes to hours[6]. In some cases, acyl-enzyme complexes can be stabilized using substrate analogues or active-site mutations but, although these approaches can provide valuable insight[7–10], they often result in complexes that are substantially non-native. Here we develop a strategy for incorporating 2,3-diaminopropionic acid (DAP) into recombinant proteins, via expansion of the genetic code[11]. We show that replacing catalytic cysteine or serine residues of enzymes with DAP permits their first-step reaction with native substrates, allowing the efficient capture of acyl-enzyme complexes that are linked through a stable amide bond. For one of these enzymes, the thioesterase domain of valinomycin synthetase[12], we elucidate the biosynthetic pathway by which it progressively oligomerizes tetradepsipeptidyl substrates to a dodecadepsipeptidyl intermediate, which it then cyclizes to produce valinomycin. By trapping the first and last acyl-thioesterase intermediates in the catalytic cycle as DAP conjugates, we provide structural insight into how conformational changes in thioesterase domains of such nonribosomal peptide synthetases control the oligomerization and cyclization of linear substrates. The encoding of DAP will facilitate the characterization of diverse acyl-enzyme complexes, and may be extended to capturing the native substrates of transiently acylated proteins of unknown function.

We proposed that selectively replacing the sulfhydryl or hydroxyl groups in catalytic cysteine or serine residues with an amino group, making 2,3-diaminopropionic acid (DAP, **1**), would enable the trapping

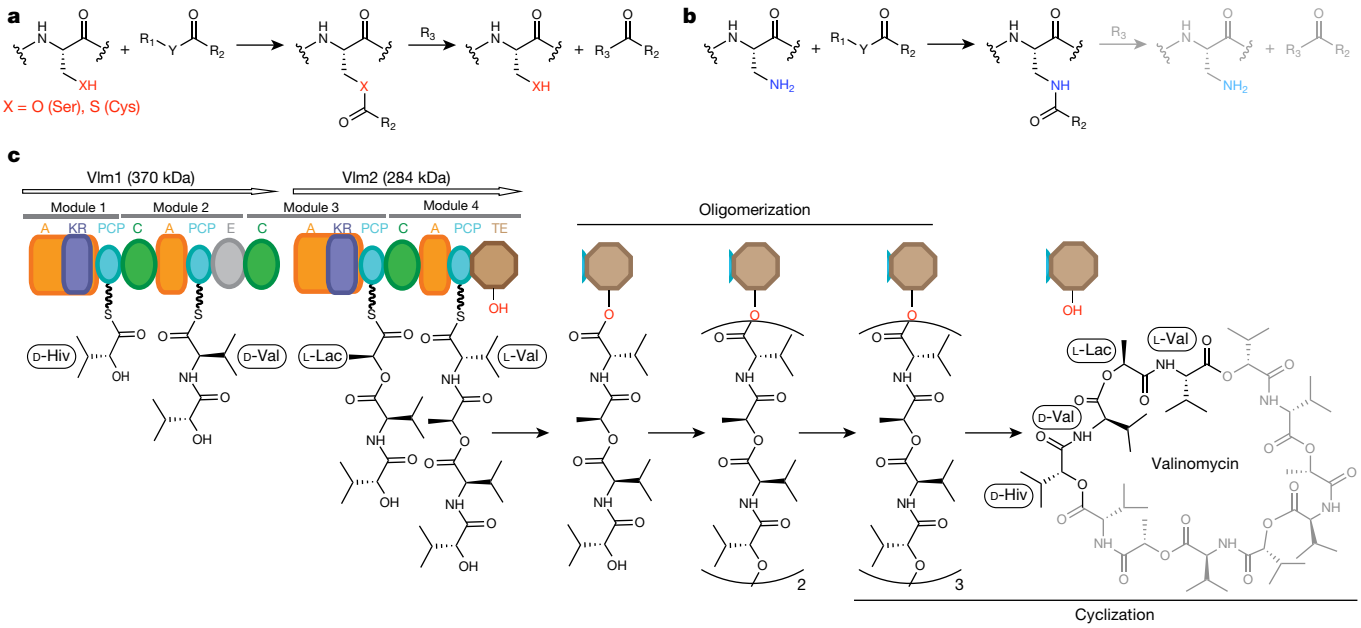


Fig. 1 | Capturing transient acyl-enzyme intermediates with DAP, and the proposed biosynthesis of valinomycin. a, Active-site serine or cysteine residues react with carbonyl groups to form tetrahedral intermediates (not shown) that collapse to acyl-enzyme intermediates by loss of R_1–YH. Attack by nucleophilic R_3 groups (commonly a hydroxyl, amine or thiol) releases the bound substrate fragment and regenerates the enzyme. R_1, R_2 and Y represent the diverse chemical groups that may be found in distinct reactants. **b**, Replacing cysteine or serine with DAP may result in a first acyl-enzyme intermediate that is resistant to cleavage. **c**, Valinomycin synthetase (Vlm) condenses D-α-hydroxyisovaleric acid (D-α-hiv), D-valine (D-val), L-lactic acid (L-lac) and L-valine (L-val) to form the tetradepsipeptidyl (D-hiv–D-val–L-lac–L-val) intermediate. D-α-hiv and L-lac arise from the reduction of precursor ketoacyl moieties by ketoreductase (KR) domains. Tetradepsipeptidyl intermediates are oligomerized to a dodecadepsipeptidyl intermediate that is cyclized, by the terminal TE domain, to produce valinomycin. Vlm1 and Vlm2 are the two protein subunits that form valinomycin synthetase. A module is a set of domains that work together to add one monomer to the growing depsipeptide. A, adenylation domain; C, condensation domain; PCP, peptidyl carrier protein domain. See Extended Data Fig. 1 for a synthetic cycle of an NRPS.

[1]Medical Research Council Laboratory of Molecular Biology, Cambridge, UK. [2]Department of Biochemistry, McGill University, Montréal, Quebec, Canada. [3]Department of Chemistry and Biomolecular Sciences, Centre for Catalysis Research and Innovation, University of Ottawa, Ottawa, Ontario, Canada. [4]These authors contributed equally: Nicolas Huguenin-Dezot, Diego A. Alonzo. *e-mail: martin.schmeing@mcgill.ca; chin@mrc-lmb.cam.ac.uk

© 2019 Springer Nature Limited. All rights reserved.

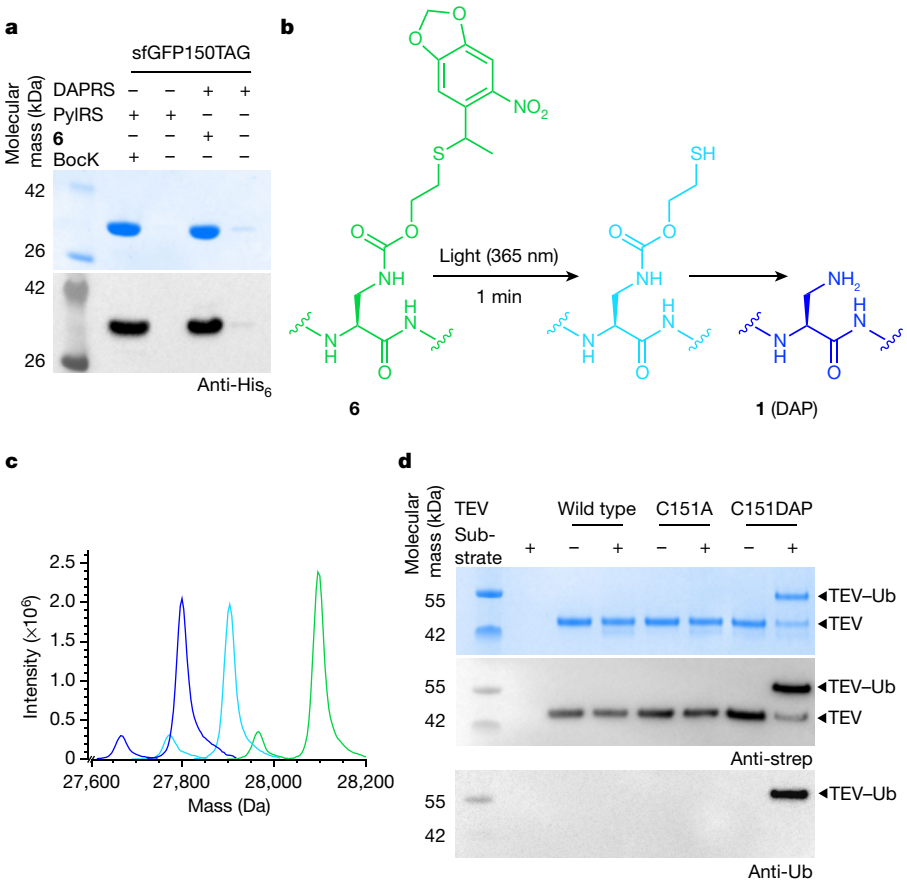


Fig. 2 | Genetically directing DAP incorporation in recombinant proteins and stably trapping the acyl-enzyme intermediate of a cysteine protease. a, SDS–PAGE gels of GFP150 (**6**) and GFP150 (BocK), with protein detected by Coomassie staining (top gel) or anti-His_6 antibody (bottom gel); the experiment was performed in two biological replicates with similar results. We used the indicated enzymes (DAPRS and PylRS) with their cognate $tRNA_{CUA}$ and amino acids (**6** and BocK (*N*ε-[(tert-butoxy)carbonyl]-L-lysine) together with an sfGFP150TAG reporter construct. **b**, Encoded **6** was photo-deprotected, leading to an intermediate, which spontaneously fragments to reveal DAP. **c**, Deprotection of **6** in sfGFP followed by electrospray ionization mass spectroscopy (ESI-MS) analysis. Green trace, purified GFP150(**6**): expected molecular mass 28,096.27 Da; observed 28,097.21 Da. Light blue trace, intermediate: expected 27,902.22 Da; observed 27,904.14 Da. Dark blue trace, incubation (10 h, 37 °C) converts the intermediate to DAP (**1**): expected 27,798.23 Da; observed 27,800.88 Da. Minor peaks resulting from loss of the N-terminal methionine are also observed. The experiment was performed in two biological replicates with similar results. **d**, TEV protease variants were incubated with Ub–tev–His. TEV(C151DAP)–Ub is the amide-bond-linked complex. Anti-Ub and anti-strep western blots confirm the identity of the complex (TEV constructs contain a streptavidin tag). The experiment was performed in two biological replicates with similar results.

of acyl-enzyme intermediates that are linked through an amide bond (Fig. 1a, b). Within peptides, the conjugate acid of the β-amino group of DAP has a reported pK_a value of between 6.3 and 7.5 (compared with the conjugate acid of the ε-amino group of free lysine, which has a pK_a of 10.5)[13]. This suggests that the β-amino group of DAP could act as a nucleophile, and may form amide bonds with the substrates of enzymes. The half-life of amides in aqueous solution is about 500 years[14], so the amide analogues of labile thioester and ester intermediates should be substantially stabilized, such that subsequent reactions with nucleophiles or solvent should be severely attenuated or abolished (Fig. 1b).

The secondary-metabolite-producing nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs) generate complex acyl-enzyme intermediates during their synthetic cycles[15]. These megaenzymes use thio-templated pathways to assemble small acyl molecules into a broad array of biologically active natural products, including antitumour compounds, antibiotics, antifungals and immunosuppressants (Extended Data Fig. 1). Unravelling their molecular mechanisms has been hampered by the challenge of characterizing their multiple acyl-enzyme intermediates at high resolution.

This challenge is exemplified by the thioesterase (TE) domains[16] from NRPS pathways that oligomerize and cyclize linear peptidyl substrates[17], including the TE domain from valinomycin synthetase[12] (Fig. 1c). This enzyme—a two-protein, four-module NRPS—alternatively links hydroxy acids (from in situ reduction of α-keto acids) and amino acids into a tetradepsipeptide intermediate, which the thioesterase domain (Vlm TE) oligomerizes up to, but not beyond, the dodecadepsipeptide. Vlm TE then cyclizes the dodecadepsipeptide to release valinomycin[12,18] (a potassium ionophore with antimicrobial, antitumoural and cytotoxic properties; Fig. 1c). The oligomerizations and cyclization must be rapid enough to prevent substantial spontaneous hydrolysis to linear depsipeptides, which are useless side products.

High-resolution structures of acyl-TE intermediates in valinomycin biosynthesis could provide mechanistic insight into how thioesterases control substrate fate. A handful of high-resolution acyl-TE structures have been obtained, most notably with the polyketide pikromycin-forming TE and non-native substrate analogues[19] (Supplementary Data 1). These have helped to identify the putative oxyanion hole and demonstrated the interaction of the 'lid' element of the TE domain with the substrate. However, structural studies of TE domains have been hampered by several factors, especially the hydrolysis rates of acyl-TE intermediates[20,21], which are high by comparison with the crystallographic timescale.

Here we develop a strategy for the site-specific incorporation of DAP into recombinant proteins, and demonstrate the efficient capture of

© 2019 Springer Nature Limited. All rights reserved.

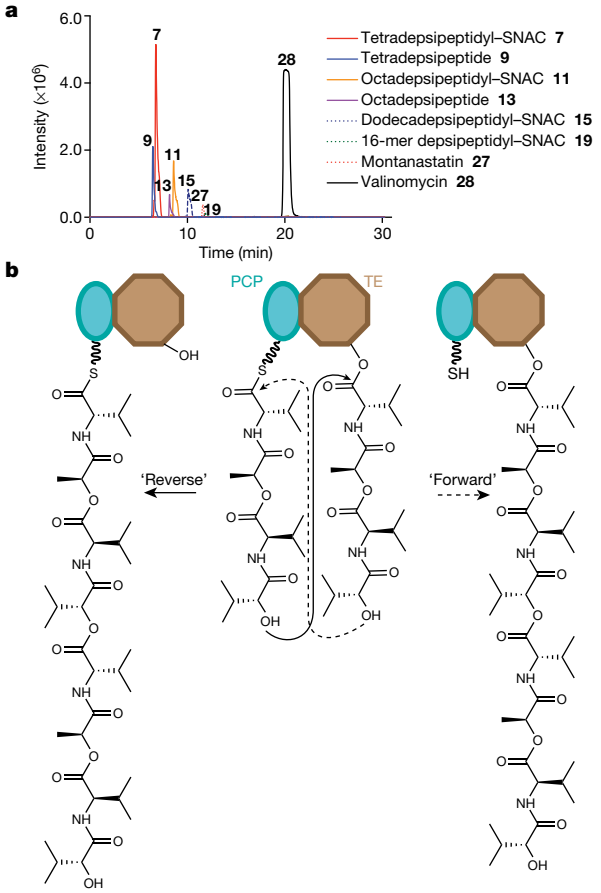


Fig. 3 | Vlm TE produces valinomycin and intermediates that delineate the oligomerization pathway from tetradepsipeptidyl–SNAC. **a**, Extracted ion chromatograms (EICs) from high-resolution (HR) liquid chromatography (LC)–ESI-MS of reactions of tetradepsipeptidyl–SNAC (**7**; 1.7 mM) and Vlm TE (6.5 μM); TE_{wt} produces valinomycin as its major product. The experiment was performed two independent times with similar results. See 'Supplementary Methods for Statistics and Reproducibility' for mass analysis and deviations from calculated *m*/*z* values. **b**, Two scenarios for oligomerization[17,26]. In the 'forward transfer' scenario, the distal hydroxyl group of tetradepsipeptidyl–*O*-TE (TE) attacks (dotted line) the thioester group in tetradepsipeptidyl–*S*-PCP (PCP), directly forming octadepsipeptidyl–*O*-TE (right). In the 'reverse' scenario, the distal hydroxyl group of tetradepsipeptidyl–*S*-PCP attacks the ester group in tetradepsipeptidyl–*O*-TE, forming octadepsipeptidyl–*S*-PCP (left), which would later be transferred onto the TE domain serine. Our data are consistent with the 'reverse' oligomerization scenario; see also Extended Data Fig. 5.

acyl-enzyme intermediates for a cysteine protease and Vlm TE. We elucidate the biosynthetic pathway for converting tetradepsipeptides to valinomycin, and structurally characterize deoxy-tetradepsipeptidyl–*N*-TE_{DAP} and dodecadepsipeptidyl–*N*-TE_{DAP} conjugates to provide insights into the first and last acyl-TE intermediates in the catalytic cycle of Vlm TE. Our results reveal how the fate of substrates may be determined by conformational changes in the TE domains of NRPSs that oligomerize and cyclize linear precursors.

The structural similarity of DAP (**1**) to cysteine and serine makes it challenging to discover an aminoacyl-tRNA synthetase that is selective for DAP in vivo. We therefore created five protected versions of DAP (**2**–**6**; Extended Data Fig. 2a), for which we anticipated that the successful discovery of specific aminoacyl-tRNA synthetase/$tRNA_{CUA}$ pairs would enable site-specific incorporation into proteins. The subsequent post-translational deprotection[22,23] would reveal DAP. We found that **2**–**5** accumulated in *Escherichia coli* at low concentrations (less than 10 μM; Extended Data Fig. 2b–e) and we were unable to evolve a synthetase for these noncanonical amino acids using several libraries of the orthogonal *Methanosarcina barkeri* (*Mb*) pyrrolysyl-tRNA synthetase (PylRS)/$tRNA^{Pyl}_{CUA}$ pair[11] (Supplementary Methods). By contrast, **6** accumulated in *E. coli* at millimolar concentrations and we were able to evolve an *Mb*PylRS variant (named DAPRS, containing mutations Y271C, N311Q, Y349F and V366C) for the site-specific incorporation of **6** (Extended Data Fig. 2f–h). The DAPRS/$tRNA^{Pyl}_{CUA}$ pair enabled the synthesis of green fluorescent protein (GFP) containing **6** at position 150 (GFP150(**6**); Fig. 2a) in good yield[24]. Photodeprotection of GFP150(**6**) and subsequent incubation converted **6** to **1** in GFP (Fig. 2b, c).

Cysteine proteases, including the tobacco etch virus (TEV) protease, react with substrates through a catalytic cysteine to generate an intermediate in which the protease is linked to the amino-terminal portion of its substrate through a thioester[4]. We replaced the active-site cysteine 151 of TEV protease with DAP, by genetically encoding **6** and deprotecting it, creating TEV(C151DAP). Incubating TEV(C151DAP) with Ub–tev–His_6, a model substrate in which the cleavage site recognized by TEV protease (tev) is flanked by ubiquitin (Ub) and a hexahistidine tag (His_6), led to cleavage of the His_6 tag from ubiquitin and formation of a covalently linked TEV(C151DAP)–Ub conjugate (Fig. 2d and Extended Data Fig. 3a–c). Control experiments demonstrated that the stable conjugate was dependent on DAP incorporation. Tandem mass spectrometry (MS/MS) demonstrated amide-bond formation between DAP and ubiquitin (Extended Data Fig. 3d). These results demonstrate that substitution of the catalytic cysteine in TEV with DAP creates a protease that performs the first step of the protease cycle, releasing the carboxy-terminal fragment of the substrate and leaving the amino-terminal fragment covalently attached to the protease through a stable amide bond that is resistant to hydrolysis.

To gain insight into the function of the TE domain and to prepare it for use with the DAP system, we expressed and purified wild-type Vlm TE (TE_{wt}). We found that Vlm TE can use an *N*-acetylcysteine (SNAC) derivative of the native depsipeptide (tetradepsipeptidyl–SNAC, **7**; Extended Data Fig. 4) to complete all stages of its catalytic cycle and yield valinomycin (Figs. 1c, 3a and Extended Data Fig. 5). Thus, **7** can mimic the natural phosphopantetheine–peptidyl carrier protein (PCP)-linked substrate, consistent with previous observations of other TE domains and substrates[17,21,25].

There are two possible pathways for the oligomerization of NRPS intermediates by TE domains, and analysis of the synthetic intermediates detected in valinomycin synthesis revealed that Vlm TE_{wt} catalyses oligomerization via a 'reverse transfer' pathway (Fig. 3b, Extended Data Fig. 5 and Supplementary Discussion 1). This pathway is analogous to that used by the more canonical gramicidin S synthetase[17,26] and we suggest that nearly all oligomerizing–cyclizing NRPSs (or PKSs[27]) will use this synthetic scheme.

We next obtained the structure of Vlm TE_{wt} (Extended Data Fig. 6 and Extended Data Table 1). It adopts the α/β-hydrolase fold typical of type I TE domains, with a canonical serine–histidine–aspartate catalytic triad[16] covered by the TE 'lid'. The lid is a mobile element with proposed roles that include substrate positioning and solvent exclusion[21,28,29]. The lid of Vlm TE is large, composed of an extended loop, three helices (Lα1–3, seen here as a bundle), a five-residue helix (Lα4), a long helix (Lα5) and another short helix (Lα6) (Extended Data Fig. 6a, b). We obtained another structure of TE_{wt} that differs only in the lid. In the first, the lid is nearly completely ordered, although the B factors are markedly higher for the Lα1–4 region, which makes almost no contact with the rest of the domain (Extended Data Fig. 6b). In the second structure, Lα4–5 have similar positions to those in the first structure, whereas Lα3 is rotated 10° towards the active site and Lα1–2 are too disordered to model.

Incubating Vlm TE with depsipeptidyl–SNACs did not yield stable conjugates (Extended Data Fig. 6c), and attempts to soak TE_{wt} crystals with depsipeptidyl–SNACs failed to reveal interpretable ligand electron density in the active site or conformational changes. Others have reported similar setbacks when attempting to visualize acyl-enzyme complexes from SNAC molecules[20,21] (Supplementary Data 1). We

© 2019 Springer Nature Limited. All rights reserved.

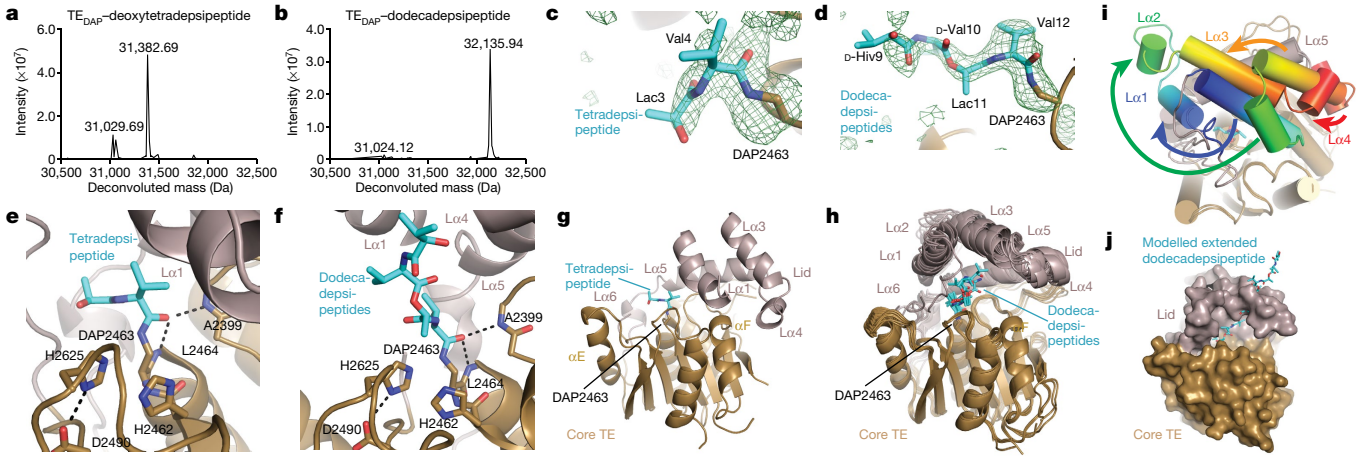


Fig. 4 | Crystal structures of complexes of TE$_{DAP}$. a, **b**, Deconvoluted mass spectra. **a**, TE$_{DAP}$ incubated with deoxy-tetradepsipeptidyl–SNAC (**8**). Expected molecular masses 31,027.24 Da (unmodified) and 31,383.70 Da (modified); observed 31,029.69 Da and 31,382.69 Da. **b**, TE$_{DAP}$ incubated with valinomycin. Expected molecular masses 31,027.24 Da (unmodified) and 32,139.11 Da (modified); observed 31,024.12 Da and 32,135.94 Da. The experiments were repeated independently five times with similar results. **c**, **d**, Unbiased electron-density ($mF_o - DF_c$) maps (green mesh, 2.5σ) for depsipeptide residues of tetradepsipeptidyl–TE$_{DAP}$ (**c**, Protein Data Bank (PDB) accession number 6ECD) and dodecadepsipeptidyl–TE$_{DAP}$ (**d**; 6ECE and 6ECF). An amide bond links DAP (brown) and depsipeptide residues (cyan). **e**, **f**, The active sites of tetradepsipeptidyl–TE$_{DAP}$ (**e**) and dodecadepsipeptidyl–TE$_{DAP}$ (**f**). The carbonyl oxygen of the amide formed by DAP and valine 4 (**e**) or valine 12 (**f**) is positioned close to the oxyanion hole formed by the main chain of A2399 and L2464. Catalytic triad residues H2625 and D2490 are shown as sticks. **g**, The lid of tetradepsipeptidyl–TE$_{DAP}$ (6ECD) is in a similar position to that seen in TE$_{wt}$ (6ECB; not shown). **h**, All crystallographically independent molecules of the dodecadepsipeptidyl–TE$_{DAP}$ (6ECE and 6ECF) are in a set of similar conformations, distinct from that seen in TE$_{wt}$. **i**, Substantial conformational changes occur in lid helices Lα1–Lα4 between the conformations of tetradepsipeptidyl–TE$_{DAP}$ (**g**) and dodecadepsipeptidyl–TE$_{DAP}$ (**h**). See also Supplementary Videos 1 and 2. **j**, In the dodecadepsipeptidyl–TE$_{DAP}$ structure, the lid sterically prevents the dodecadepsipeptide from extending out in a linear fashion, instead favouring it curling back through this steric block and forming largely hydrophobic, non-specific interactions with the lid.

conclude that acyl intermediates in valinomycin biosynthesis are not stable, and that it is exceptionally challenging to use wild-type Vlm TE to visualize biosynthetic intermediates.

We therefore produced Vlm TE in which the active-site serine 2463 was replaced by DAP (TE$_{DAP}$; Extended Data Fig. 7a–c) in order to capture stable acyl-TE conjugates (Fig. 4). To provide insight into the first acyl-thioesterase intermediate in the catalytic cycle of Vlm TE, we captured tetradepsipeptidyl–*N*-TE$_{DAP}$: incubation of TE$_{DAP}$ with tetradepsipeptidyl–SNAC (**7**) led to production of a stable depsipeptidyl–TE$_{DAP}$ intermediate in greater than 60% yield (Extended Data Fig. 7d), and we did not observe valinomycin synthesis. However, remarkably, we did observe a small amount of octadepsipeptidyl–SNAC (**11**; Extended Data Fig. 5f, h). **11** is probably formed by enzyme-catalysed attack of the tetradepsipeptidyl–SNAC's hydroxyl group on the amide bond that links the tetradepsipeptide to TE$_{DAP}$. The attack of a hydroxyl on an amide is analogous to the first reaction used by related serine proteases[30], but it is surprising that this TE domain is capable of catalysing a more demanding chemical reaction (amide cleavage) than the reaction (ester hydrolysis) it evolved to perform. In an effort to enhance conjugate yield, we optimized the conditions for conjugating deoxy-tetradepsipeptidyl–SNAC **8** with TE$_{DAP}$, which produced (in about 70% yield) the deoxy-depsipeptidyl–TE$_{DAP}$ conjugate (Fig. 4a). The marginal solubility of these hydrophobic SNACs may limit the conjugation efficiency.

To determine the structure of the deoxy-tetradepsipeptidyl–*N*-TE$_{DAP}$ conjugate, we incubated TE$_{DAP}$ crystals with the deoxy-tetradepsipeptidyl–SNAC (**8**). The resulting electron density shows somewhat weak but unambiguous density for an amide bond between DAP 2463 and L-valine 4 of the deoxy-tetradepsipeptide (Fig. 4c and Extended Data Fig. 8a). The carbonyl oxygen of the L-valine 4 is close to backbone amides of residues alanine 2399 and leucine 2464—the putative oxyanion hole[25] (Fig. 4e). There is also density for the next residue, L-lactic acid 3 (L-lac3), but it is insufficient to reliably model D-valine 2 and D-α-hydroxyisovaleric acid 1 (D-hiv1) as the deoxy-tetradepsipeptide arcs out, indicating substrate flexibility. The deoxy-tetradepsipeptide does not make any interactions with the lid, which is in a conformation nearly identical to that in the first TE$_{wt}$ (apo) structure (Fig. 4g and Extended Data Fig. 6d).

Next, we sought insight into the last acyl-TE intermediate in the catalytic cycle. Upon incubation of valinomycin and TE$_{DAP}$, we captured dodecadepsipeptidyl–*N*-TE$_{DAP}$ in 65–100% yield, formed through a ring-opening reaction analogous to the reverse of the natural cyclization (Fig. 4b). This reaction is thermodynamically favoured by virtue of amide-bond formation. Dodecadepsipeptidyl–TE$_{DAP}$ produced crystals in similar conditions to those of TE$_{wt}$, but with a different morphology and belonging to two different space groups (*H*3 and *P*1, with two and six molecules per asymmetric unit, respectively; Extended Data Table 1). All eight crystallographically independent molecules of dodecadepsipeptidyl–TE$_{DAP}$ showed some density for the dodecadepsipeptide. Molecules P1_A–F and H3_A–B show strong density for four, three, two, two, two, two, three and one dodecadepsipeptide residues respectively (Fig. 4d and Extended Data Fig. 8b–i). Additional weaker density is present in some molecules, which could accommodate up to the full 12 residues (Extended Data Fig. 8j–l); in others, weaker density suggests multiple conformations for the distal residues, but they were not possible to definitively model. The modelled depsipeptides all follow a similar trajectory away from the active site. There is no consistent interaction between the depsipeptide beyond the L-valine residue attached to DAP and the TE domain (Fig. 4f, h). Rather, each depsipeptide makes different contacts with the lid. The lid forms a semi-sphere-like pocket/steric barrier made up of helices Lα1, 3, 4 and 5, and the strand amino-terminal to Lα1. The lid of each crystallographically independent molecule of dodecadepsipeptidyl–TE$_{DAP}$ is in a similar but non-identical position, and the loops between lid helices are disordered in most molecules (Fig. 4h). This again highlights the mobility of the lid and explains why the conformation and extent of order of dodecadepsipeptides differ between molecules (Fig. 4h). The semi-sphere-like barrier occurs only because of a major rearrangement of the lid in the dodecadepsipeptidyl–TE$_{DAP}$ structures with respect to the conformation of the lid seen in both the apo and tetradepsipeptidyl-bound structures of Vlm TE.

© 2019 Springer Nature Limited. All rights reserved.

Comparing the position of the Vlm TE lid in the apo and tetradepsipeptide-bound structures with the position of the lid in the dodecadepsipeptide-bound structures demonstrates and emphasizes its extreme mobility. To transition from one lid conformation to the other, helices Lα5–6 maintain their position, Lα3–4 rotate by about 45° and translocate roughly 13 Å, Lα2 translocates roughly 25 Å, and Lα1 shortens, translocates about 13 Å and rotates more than 90° in the opposite direction to Lα3–4 (Fig. 4i and Supplementary Videos 1, 2). This dramatic rearrangement means that the lid helices pack together in a markedly different manner in the apo/tetradepsipeptidyl-bound structure and in dodecadepsipeptidyl-bound conformations.

The distinct lid conformations directly influence the possible location of the depsipeptide. In the apo/tetradepsipeptide-bound conformation of the lid, the carboxyl terminus of Lα1 comes within 10 Å of serine/DAP 2463, leading the tetradepsipeptide to extend towards the TE core helix αE. In the dodecadepsipeptide-bound conformations of the lid, the loop adjacent to Lα1 blocks the location occupied by the tetradepsipeptide in the tetradepsipeptide-bound structure. Moreover, in the dodecadepsipeptide-bound structure the amino terminus of Lα1 forms part of the semi-sphere-like pocket. This pocket probably helps to curl the dodecadepsipeptide back towards serine/DAP 2463, entropically controlling cyclization as part of the oligomerization/cyclization pathway (Fig. 4j, Extended Data Fig. 9 and Supplementary Discussion 2).

In summary, we have genetically encoded DAP in place of catalytic cysteine and serine residues to capture unstable thioester or ester intermediates as stable amide analogues. We have exemplified the utility of this approach for a cysteine protease and a thioesterase, and provided unique insight into intermediates in the synthesis of valinomycin: a massive lid rearrangement that is associated with the dodecadepsipeptidyl-bound Vlm TE reorients the substrate from its position during oligomerization and places it into a pocket that entropically controls cyclization. Importantly, the DAP system enables the formation of near-native acyl-enzyme complexes with widely used, reaction-competent substrates (for example, native proteins containing protease sites), substrate analogues (here SNACs), and commercially available natural products (here valinomycin, and probably other cyclic products[25]). We anticipate that the approach will be broadly applicable and may be extended to capturing native substrates of transiently acylated proteins of unknown function.

Reporting summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Source data for all figures are available from the corresponding authors upon reasonable request. The models and structure factors for the crystal structures are deposited in the Protein Data Bank with accession numbers 6ECB, 6ECC, 6ECD, 6ECE and 6ECF. Detailed methods, including chemical syntheses, are available in the Supplementary Information.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0781-z.

Received: 16 August 2018; Accepted: 2 November 2018;
Published online 12 December 2018.

1. Holliday, G. L., Mitchell, J. B. O. & Thornton, J. M. Understanding the functional roles of amino acid residues in enzyme catalysis. *J. Mol. Biol.* **390**, 560–577 (2009).
2. Hedstrom, L. Serine protease mechanism and specificity. *Chem. Rev.* **102**, 4501–4524 (2002).
3. Long, J. Z. & Cravatt, B. F. The metabolic serine hydrolases and their functions in mammalian physiology and disease. *Chem. Rev.* **111**, 6022–6063 (2011).
4. Otto, H. H. & Schirmeister, T. Cysteine proteases and their inhibitors. *Chem. Rev.* **97**, 133–172 (1997).
5. Swatek, K. N. & Komander, D. Ubiquitin modifications. *Cell Res.* **26**, 399–422 (2016).
6. Yang, W. & Drueckhammer, D. G. Understanding the relative acyl-transfer reactivity of oxoesters and thioesters: computational analysis of transition state delocalization effects. *J. Am. Chem. Soc.* **123**, 11004–11009 (2001).
7. Liu, B., Schofield, C. J. & Wilmouth, R. C. Structural analyses on intermediates in serine protease catalysis. *J. Biol. Chem.* **281**, 24024–24035 (2006).
8. Scaglione, J. B. et al. Biochemical and structural characterization of the tautomycetin thioesterase: analysis of a stereoselective polyketide hydrolase. *Angew. Chem. Int. Ed.* **49**, 5726–5730 (2010).
9. Cappadocia, L. & Lima, C. D. Ubiquitin-like protein conjugation: structures, chemistry, and mechanism. *Chem. Rev.* **118**, 889–918 (2018).
10. Plechanovová, A., Jaffray, E. G., Tatham, M. H., Naismith, J. H. & Hay, R. T. Structure of a RING E3 ligase and ubiquitin-loaded E2 primed for catalysis. *Nature* **489**, 115–120 (2012).
11. Chin, J. W. Expanding and reprogramming the genetic code of cells and animals. *Annu. Rev. Biochem.* **83**, 379–408 (2014).
12. Magarvey, N. A., Ehling-Schulz, M. & Walsh, C. T. Characterization of the cereulide NRPS alpha-hydroxy acid specifying modules: activation of alpha-keto acids and chiral reduction on the assembly line. *J. Am. Chem. Soc.* **128**, 10698–10699 (2006).
13. Lan, Y. et al. Incorporation of 2,3-diaminopropionic acid into linear cationic amphipathic peptides produces pH-sensitive vectors. *ChemBioChem* **11**, 1266–1272 (2010).
14. Radzicka, A. & Wolfenden, R. Rates of uncatalyzed peptide bond hydrolysis in neutral solution and the transition state affinities of proteases. *J. Am. Chem. Soc.* **118**, 6105–6109 (1996).
15. Reimer, J. M., Haque, A. S., Tarry, M. J. & Schmeing, T. M. Piecing together nonribosomal peptide synthesis. *Curr. Opin. Struct. Biol.* **49**, 104–113 (2018).
16. Horsman, M. E., Hari, T. P. A. & Boddy, C. N. Polyketide synthase and non-ribosomal peptide synthetase thioesterase selectivity: logic gate or a victim of fate? *Nat. Prod. Rep.* **33**, 183–202 (2016).
17. Hoyer, K. M., Mahlert, C. & Marahiel, M. A. The iterative gramicidin s thioesterase catalyzes peptide ligation and cyclization. *Chem. Biol.* **14**, 13–22 (2007).
18. Jaitzig, J., Li, J., Süssmuth, R. D. & Neubauer, P. Reconstituted biosynthesis of the nonribosomal macrolactone antibiotic valinomycin in *Escherichia coli*. *ACS Synth. Biol.* **3**, 432–438 (2014).
19. Akey, D. L. et al. Structural basis for macrolactonization by the pikromycin thioesterase. *Nat. Chem. Biol.* **2**, 537–542 (2006).
20. Bruner, S. D. et al. Structural basis for the cyclization of the lipopeptide antibiotic surfactin by the thioesterase domain SrfTE. *Structure* **10**, 301–310 (2002).
21. Samel, S. A., Wagner, B., Marahiel, M. A. & Essen, L. O. The thioesterase domain of the fengycin biosynthesis cluster: a structural base for the macrocyclization of a non-ribosomal lipopeptide. *J. Mol. Biol.* **359**, 876–889 (2006).
22. Li, J. et al. Palladium-triggered deprotection chemistry for protein activation in living cells. *Nat. Chem.* **6**, 352–361 (2014).
23. Baker, A. S. & Deiters, A. Optical control of protein function through unnatural amino acid mutagenesis and other optogenetic approaches. *ACS Chem. Biol.* **9**, 1398–1407 (2014).
24. Virdee, S., Ye, Y., Nguyen, D. P., Komander, D. & Chin, J. W. Engineered diubiquitin synthesis reveals Lys29-isopeptide specificity of an OTU deubiquitinase. *Nat. Chem. Biol.* **6**, 750–757 (2010).
25. Tseng, C. C. et al. Characterization of the surfactin synthetase C-terminal thioesterase domain as a cyclic depsipeptide synthase. *Biochemistry* **41**, 13350–13359 (2002).
26. Trauger, J. W., Kohli, R. M., Mootz, H. D., Marahiel, M. A. & Walsh, C. T. Peptide cyclization catalysed by the thioesterase domain of tyrocidine synthetase. *Nature* **407**, 215–218 (2000).
27. Zhou, Y., Prediger, P., Dias, L. C., Murphy, A. C. & Leadlay, P. F. Macrodiolide formation by the thioesterase of a modular polyketide synthase. *Angew. Chem. Int. Ed.* **54**, 5232–5235 (2015).
28. Frueh, D. P. et al. Dynamic thiolation-thioesterase structure of a non-ribosomal peptide synthetase. *Nature* **454**, 903–906 (2008).
29. Whicher, J. R. et al. Structure and function of the RedJ protein, a thioesterase from the prodiginine biosynthetic pathway in *Streptomyces coelicolor*. *J. Biol. Chem.* **286**, 22558–22569 (2011).
30. Ekici, O. D., Paetzel, M. & Dalbey, R. E. Unconventional serine proteases: variations on the catalytic Ser/His/Asp triad configuration. *Protein Sci.* **17**, 2023–2037 (2008).

Acknowledgements We thank P. Emsley and G. Murshudov for their help in modelling depsipeptides; A. Wahba, J. Reimer, G. Bridon, K. Heesom and T. Elliott for help with mass spectrometry; M. Tarry and N. Rogerson for editing; C. Alonso for early crystal trials; R. Hay for early discussions on DAP; and S. Zhang for help with cloning. We also thank the staff of beamlines CLS 08ID-1 (S. Labiuk, J. Gorin, M. Fodje, K. Janzen, D. Spasyuk and P. Grochulski) and APS 24-ID-C (grants GM124165, RR029205, DE-AC02-06CH11357; F. Murphy). This work is supported by grants to J.W.C. from the Medical Research Council, UK (grants MC_U105181009 and MC_UP_A024_1008), to T.M.S. from the Canada Research Chair and the Natural Sciences and Engineering Research Council of Canada (NSERC; Discovery Grant 418420) and to C.N.B. from the NSERC (Discovery Grant 06167).

Author contributions N.H.-D. and D.A.A. contributed equally. M.M and G.W.H. contributed equally. N.H.-D. and D.P.N. selected the synthetases for **2**–**6**. N.H.-D.

© 2019 Springer Nature Limited. All rights reserved.

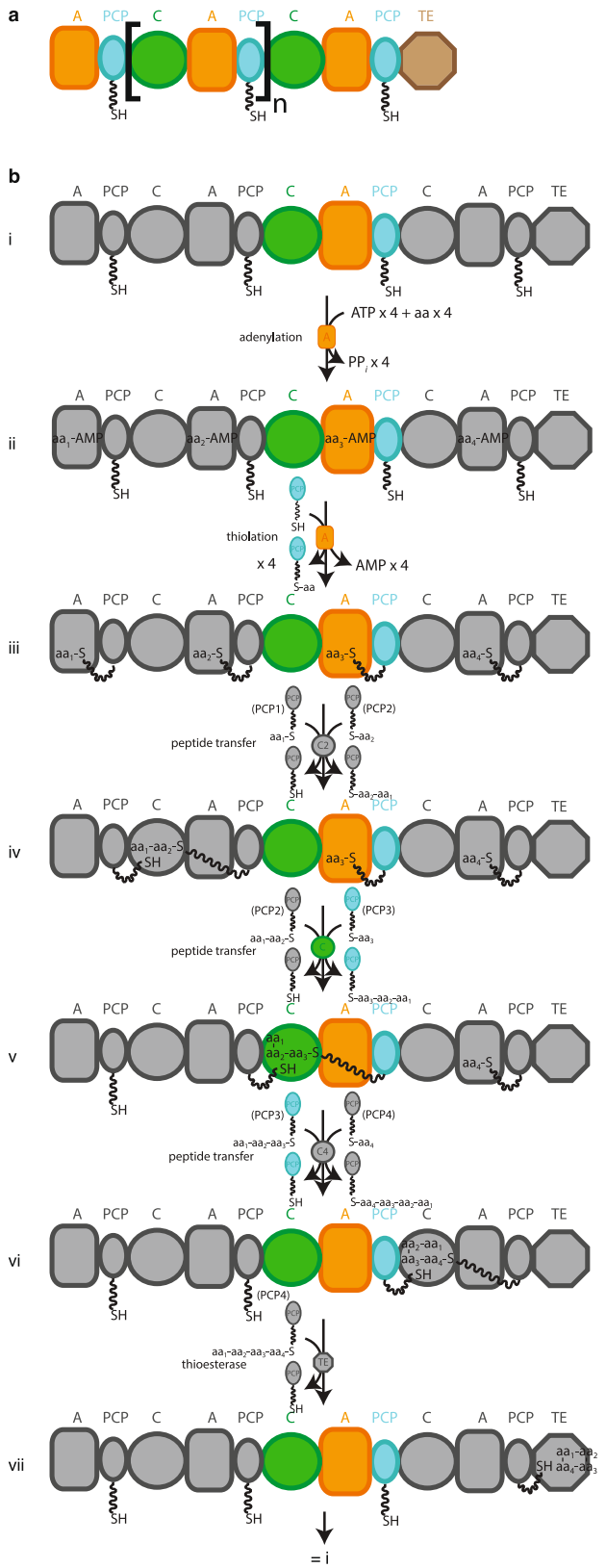
characterized the incorporation of **6** and its deprotection to produce DAP, and performed the TEV conjugation. D.P.N and M.M. designed, and M.M. synthesized **2**–**6**. T.M.S. and D.A.A. designed, and D.A.A. performed, the biochemistry and crystallography with Vlm TE_{wt}. D.A.A. and N.H.-D. performed biochemistry and crystallography with Vlm TE_{DAP}. G.W.H. and M.H.D. synthesized Vlm TE substrates. J.W.C. supervised N.H.-D., D.P.N. and M.M. T.M.S. supervised D.A.A. C.N.B. supervised G.W.H. and M.H.D. T.M.S., D.A.A., N.H.-D. and J.W.C. wrote the paper with input from all authors.

Competing interests The authors declare no competing interests.

Additional information

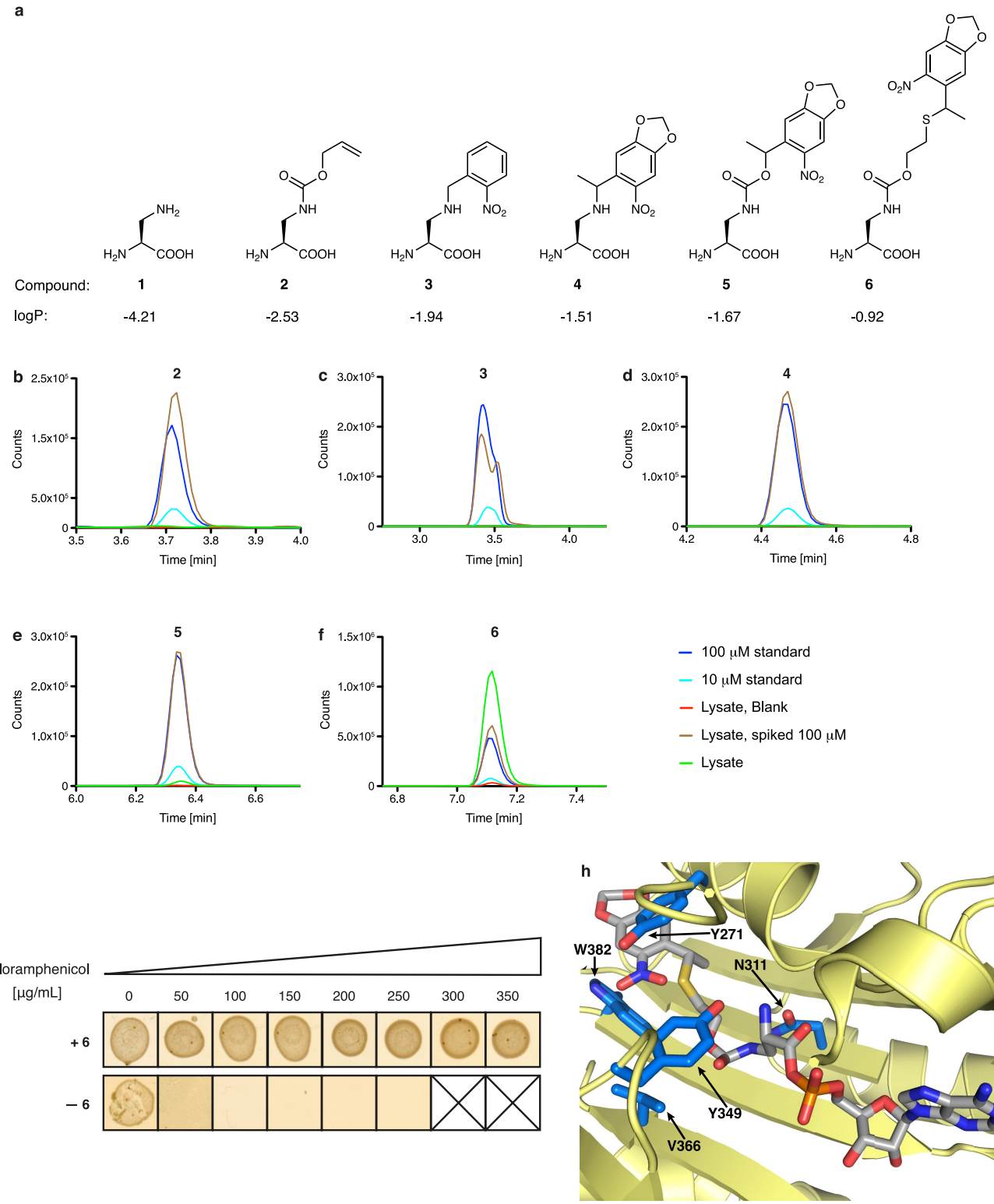
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0781-z.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0781-z.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to T.M.S. or J.W.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 1 | Schematic representation and reaction cycle of a canonical NRPS. a, Schematic representation of a generic type I NRPS. The square brackets denote a single module. **b**, i–vii, Synthetic cycle of a canonical elongation module. NRPSs assemble peptides from amino acyl and other small acyl building blocks using a modular and thio-templated logic. A canonical NRPS is composed of one module for every residue in the peptide product. The initiation module contains an adenylation (A) domain, which binds cognate acyl substrate and performs adenylation and transfer of that substrate as a thioester on the phosphopantetheine arm (PPE, shown as a wavy line) of a peptidyl carrier protein (PCP) domain, for transport between active sites. Each elongation module contains an A and a PCP domain, and also a condensation (C) domain, which condenses aminoacyl and peptidyl substrates bound to PCP domains, thus progressively elongating the nascent chain. Termination modules contain C, A and PCP domains, and a specialized terminating/offloading domain responsible for the release of the peptide in its final form. The most common and most versatile terminating domain in NRPSs is the TE domain. Similar TE domains terminate synthesis in polyketide and fatty acid synthases. PP_i, diphosphate; aa, amino acid.

© 2019 Springer Nature Limited. All rights reserved.



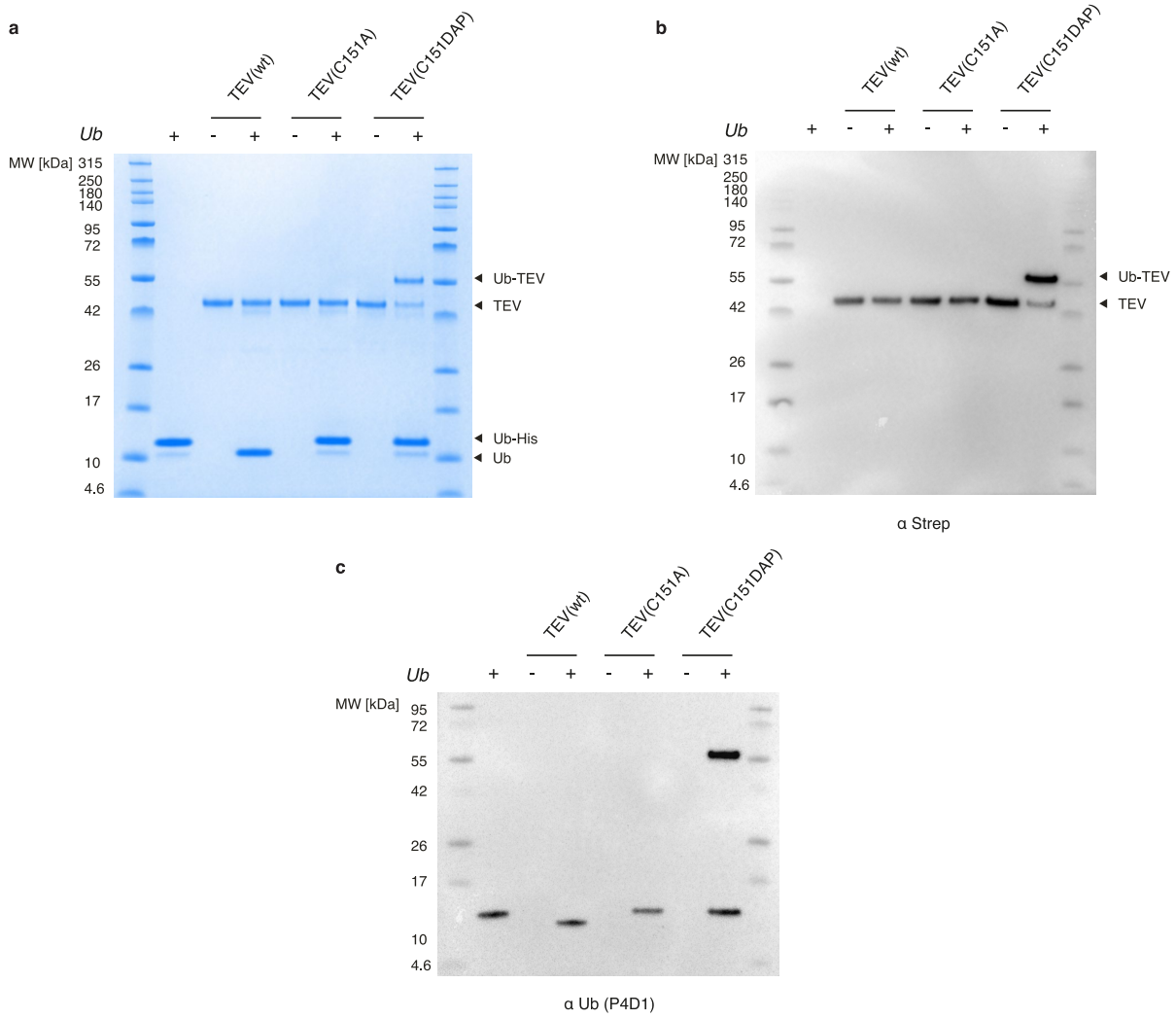
Extended Data Fig. 2 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Fig. 2 | Genetically directing DAP incorporation in recombinant proteins. a, Structure of DAP and the protected versions investigated herein. **1**, 2,3-diaminopropionic acid (DAP); **2**, (*S*)-3-(((allyloxy)carbonyl)amino)-2-aminopropanoic acid; **3**, (*S*)-2-amino-3-((2-nitrobenzyl)amino)propanoic acid; **4**, (2*S*)-2-amino-3-((1-(6-nitrobenzo[*d*][1,3]dioxol-5-yl)ethyl)amino)propanoic acid; **5**, (2*S*)-2-amino-3-(((1-(6-nitrobenzo[*d*][1,3]dioxol-5-yl)ethoxy)carbonyl)amino)propanoic acid; **6**, (2*S*)-2-amino-3-(((2-((1-(6-nitrobenzo[*d*][1,3]dioxol-5-yl)ethyl)thio)ethoxy)carbonyl)amino)propanoic acid. Calculated log*P* values are indicated (calculated using the Molinspiration molecular property calculation services at www.molinspiration.com/cgi-bin/properties). **b**–**f**, Determining the intracellular concentration of compounds **2**–**6** by an LC–MS assay, performed on extracts. The dark-blue trace represents a 100 μM standard for each compound. The light-blue trace represents a 10 μM standard for each compound. The red trace results from cells grown in the absence of the compound. The brown trace results from cells grown in the absence of the compound, but spiked with the compound to 100 μM. The green trace results from cells grown in the presence of 1 mM compound. The experiments were repeated in two biological replicates with similar results. **g**, Phenotyping of the DAPRS/$tRNA_{CUA}$ pair. Cells containing the DAPRS/$tRNA_{CUA}$ pair and *cat*(*112TAG*) (encoding a chloramphenicol-resistance gene containing an amber stop codon (TAG) at codon 112) were plated in the presence or absence of **6** on the indicated concentrations of chloramphenicol. The experiment was performed in two biological replicates with similar results. **h**, The side chain of **6** (grey sticks) was modelled into the active site of PylRS using a co-crystal structure of PylRS and adenylated pyrrolysine (PDB accession number 2ZIM[31]). PylRS is displayed in pale yellow and amino-acid positions randomized in DAPRSlib are shown in marine blue.

31. Kavran, J. M. et al. Structure of pyrrolysyl-tRNA synthetase, an archaeal enzyme for genetic code innovation. *Proc. Natl Acad. Sci. USA* **104**, 11268–11273 (2007).

© 2019 Springer Nature Limited. All rights reserved.

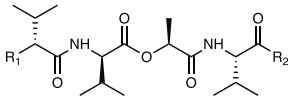


Extended Data Fig. 3 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Fig. 3 | Stably trapping the acyl-enzyme intermediate of a cysteine protease. a, Different variants of TEV protease (shown at the top) were reacted with Ub–tev–His. The use of TEV(wt) results in cleavage of the TEV cleavage sequence. The use of TEV(C151A) results in minimal cleavage. The presence of DAP in the active site of TEV results in the presence of an extra band in the Coomassie gel, representing the isopeptide-linked TEV(C151DAP)–Ub complex. **b**, **c**, Anti-streptavidin (α-strep; **b**) and anti-Ub (α-Ub antibody P4D1; **c**) western blots of the reactions confirm the identity of the complex. For **a**–**c**, the experiment was repeated in two biological replicates with similar results. **d**, Tandem mass spectrometry following tryptic digest of the TEV(C151DAP)–Ub conjugate confirms amide-bond formation at the expected position. Top, the sequence of the branched peptide subject to fragmentation. Fragmentation of the substrate chain is predicted to lead to a series of y ions (yellow) and a series of b ions (green); the ions from this chain are labelled as 'β'. Fragmentation of the TEV(C151DAP)-derived chain is predicted to lead to a series of y ions (blue) and a series of b ions (red); the ions from this chain are labelled as 'α'. Bottom, MS/MS spectra with peak assignments. Ions in the α-chain were assigned by treating DAP and the β-chain as a modification of known mass. Ions in the β-chain were manually assigned. The mass-spectrometry analysis was performed once.

© 2019 Springer Nature Limited. All rights reserved.

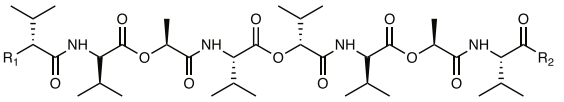


7 R_1=OH, R_2=$SCH_2CH_2NHC(O)CH_3$

8 R_1=H, R_2=$SCH_2CH_2NHC(O)CH_3$

9 R_1=OH, R_2=OH

10 R_1=H, R_2=OH

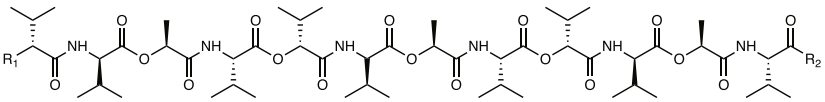


11 R_1=OH, R_2=$SCH_2CH_2NHC(O)CH_3$

12 R_1=H, R_2=$SCH_2CH_2NHC(O)CH_3$

13 R_1=OH, R_2=OH

14 R_1=H, R_2=OH

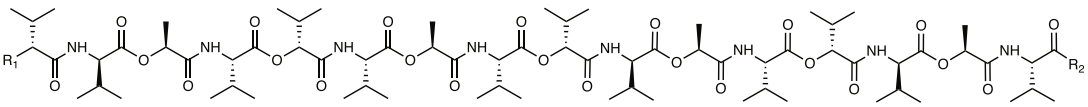


15 R_1=OH, R_2=$SCH_2CH_2NHC(O)CH_3$

16 R_1=H, R_2=$SCH_2CH_2NHC(O)CH_3$

17 R_1=OH, R_2=OH

18 R_1=H, R_2=OH

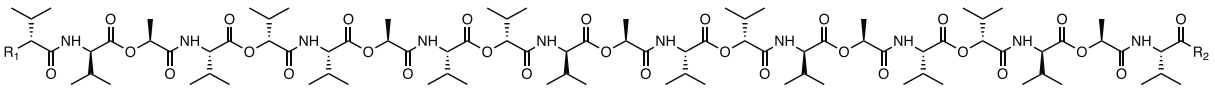


19 R_1=OH, R_2=$SCH_2CH_2NHC(O)CH_3$

20 R_1=H, R_2=$SCH_2CH_2NHC(O)CH_3$

21 R_1=OH, R_2=$SCH_2CH_2NHC(O)CH_3$

22 R_1=H, R_2=OH

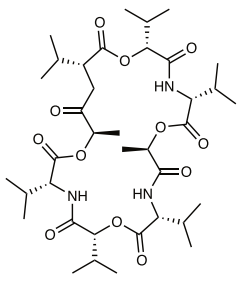


23 R_1=OH, R_2=$SCH_2CH_2NHC(O)CH_3$

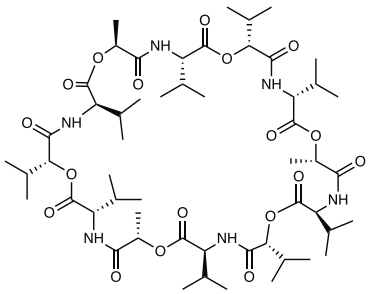
24 R_1=H, R_2=$SCH_2CH_2NHC(O)CH_3$

25 R_1=OH, R_2=OH

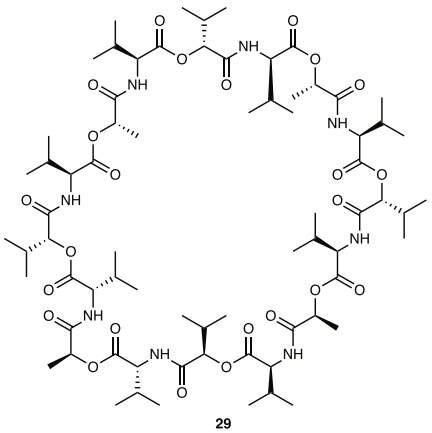
26 R_1=H, R_2=OH



montanastatin (cyclic octadepsipeptide), **27**



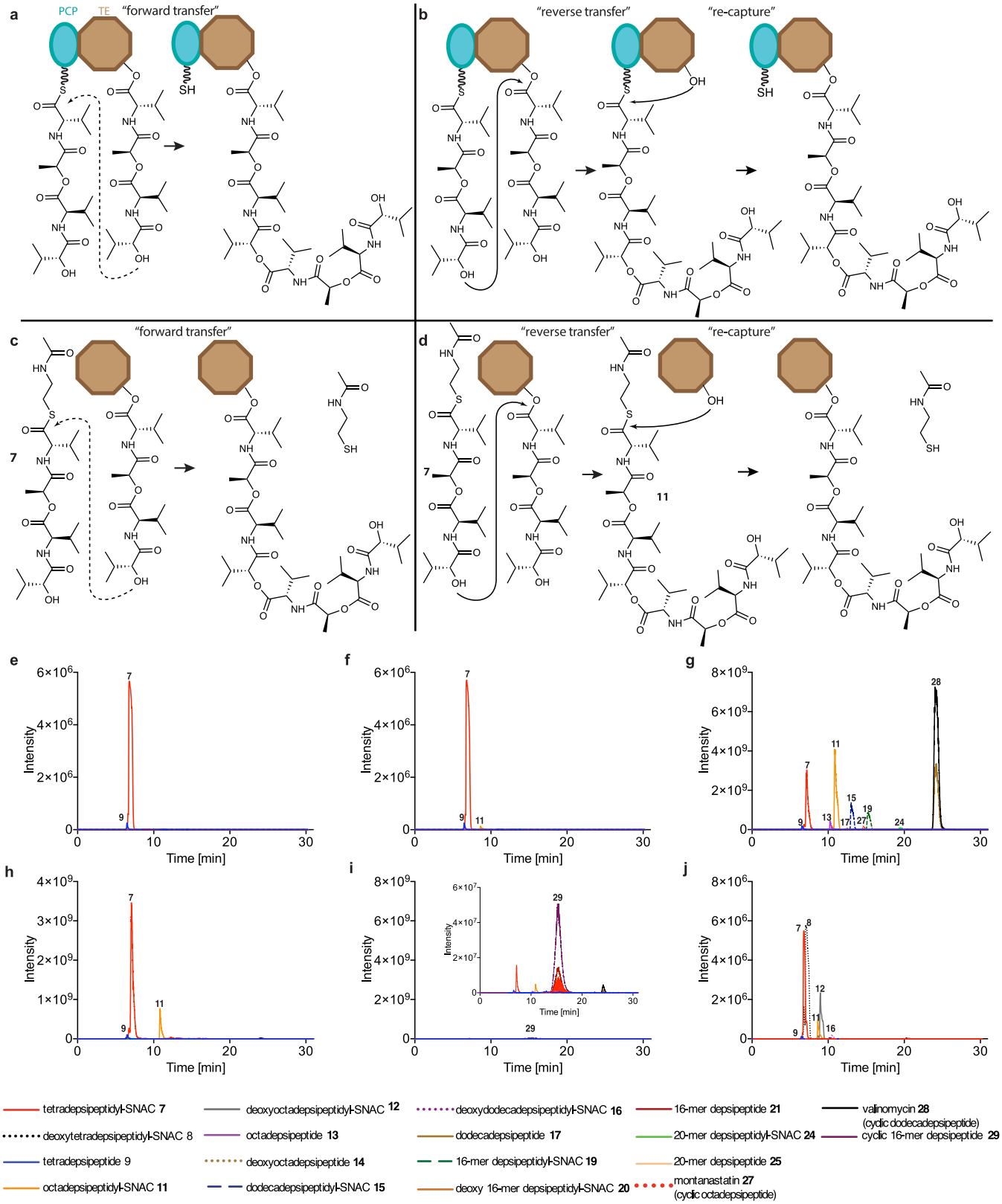
valinomycin, **28**



29

Extended Data Fig. 4 | Chemical structures of key Vlm TE substrates and products. The chemical structures and the numbers used to refer to them are shown.

© 2019 Springer Nature Limited. All rights reserved.

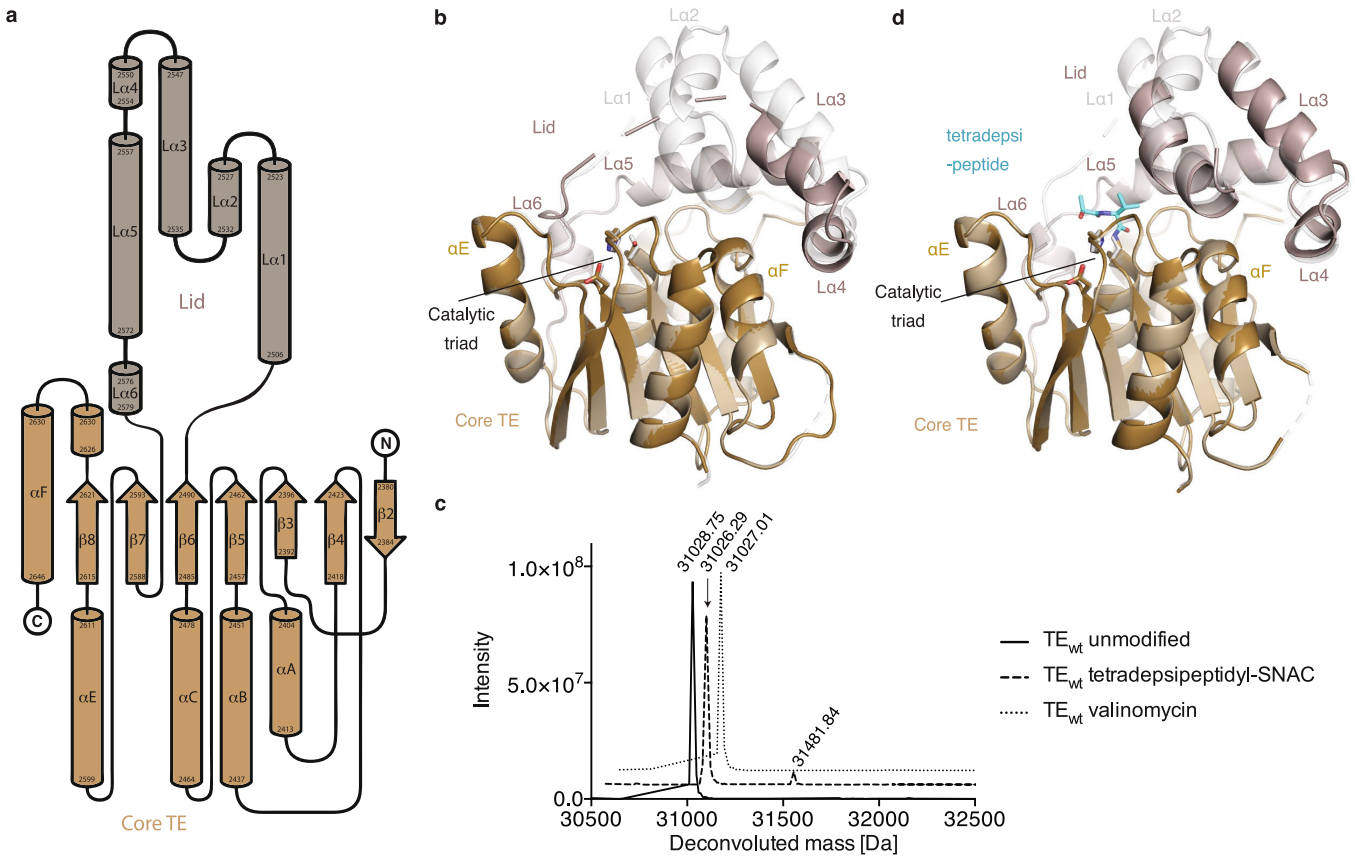


Extended Data Fig. 5 | See next page for caption.

© 2019 Springer Nature Limited. All rights reserved.

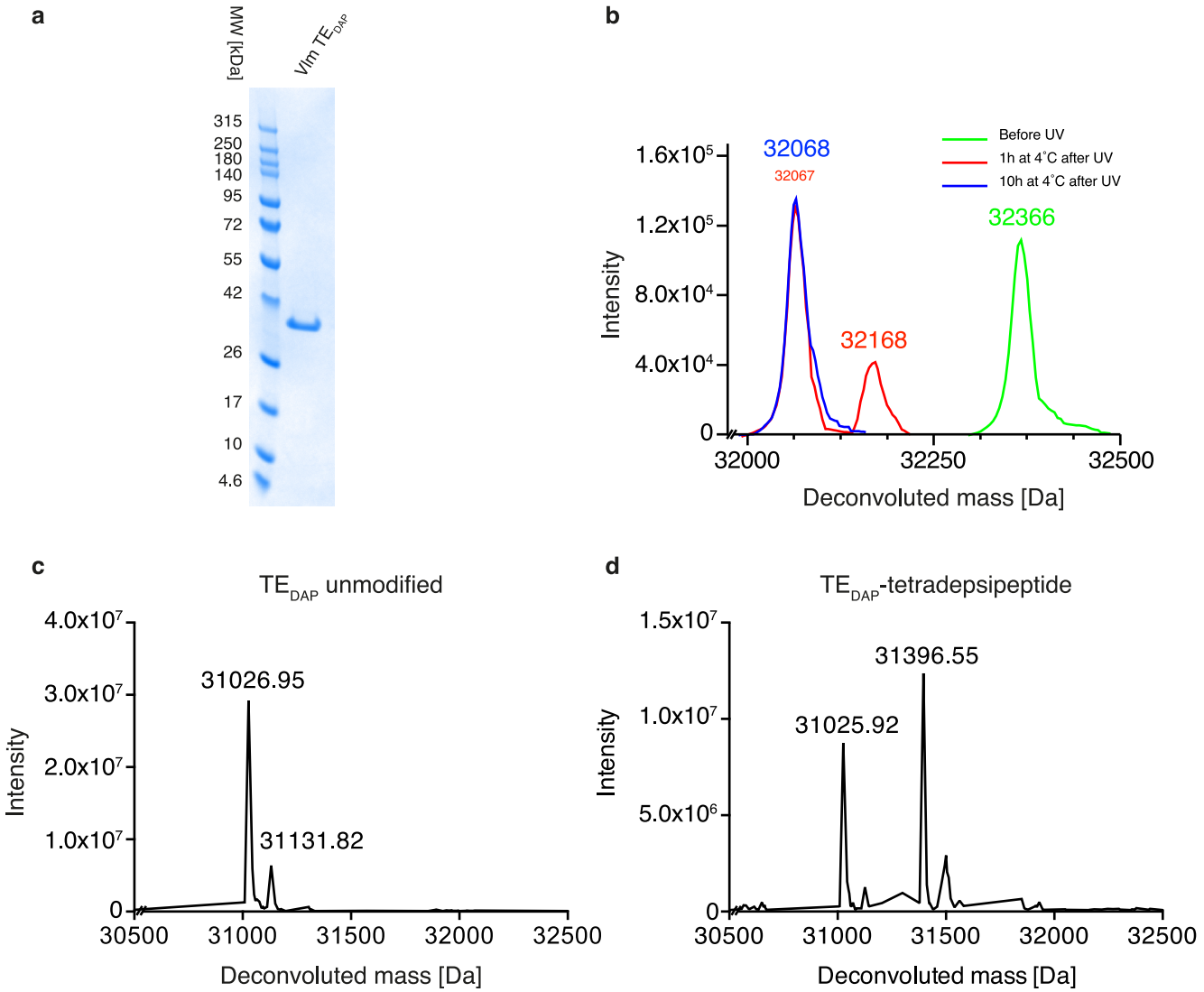
Extended Data Fig. 5 | The mechanism by which by Vlm TE catalyses oligomerization. Oligomerization could conceivably take place in two ways. **a**, In the first scenario, 'forward transfer', the distal hydroxyl group of the tetradepsipeptidyl–*O*-TE complex attacks the thioester group in the tetradepsipeptidyl–*S*-PCP enzyme intermediate, directly forming octadepsipeptidyl–*O*-TE as a product. **b**, In the second scenario, 'reverse transfer', the distal hydroxyl group of the tetradepsipeptidyl–*S*-PCP complex attacks the ester group in the tetradepsipeptidyl–*O*-TE enzyme intermediate, forming octadepsipeptidyl–*S*-PCP as a product, which would then need to be transferred onto the TE-domain serine (here labelled as 're-capture'). **c**, **d**, Analogous scenarios involving tetradepsipeptidyl–SNAC (**7**) as the substrate instead of tetradepsipeptidyl–*S*-PCP. **e**, **f**, EICs (HR LC–ESI-MS) of a mix of **7** (1.7 mM) and buffer (**e**), or the products of a reaction between **7** (1.7 mM) and Vlm TE_{DAP} (6.5 μM) (**f**). **g**–**i**, EICs (low-resolution (LR) LC–ESI-MS) of reactions using a higher-volume injection into an ion-trap MS instrument. **g**, The higher-volume injection of a reaction of **7** (1.7 mM) and Vlm TE_{wt} (6.5 μM) enabled detection of a peak consistent with the 20-mer depsipeptidyl–SNAC (**24**). **h**, LC ion-trap MS of the reaction of **7** (1.7 mM) and Vlm TE_{DAP} (6.5 μM). **i**, Small amounts of the cyclic 16-mer depsipeptide **29** elute during post-run column clean-up of experiment shown in **g**. **j**, EICs (HR LC–ESI-MS) of products of reactions between Vlm TE_{wt} (6.5 μM) and a mix of **7** and deoxy-tetradepsipeptidyl–SNAC (**8**; 1.7 mM of each). TE_{wt} produces the intermediates deoxy-octadepsipeptidyl–SNAC (**12**), deoxy-dodecadepsipeptidyl–SNAC (**16**) and deoxy 16-mer depsipeptidyl–SNAC (**20**), confirming the reaction pathway shown in **b**. See 'Supplementary Methods for Statistics and Reproducibility' for accurate mass analysis and deviations from calculated *m*/*z* values of each compound. The experiments in **e**–**i** were repeated independently twice with similar results. Mass-spectrometry analysis of the experiment in **j** was performed once.

© 2019 Springer Nature Limited. All rights reserved.



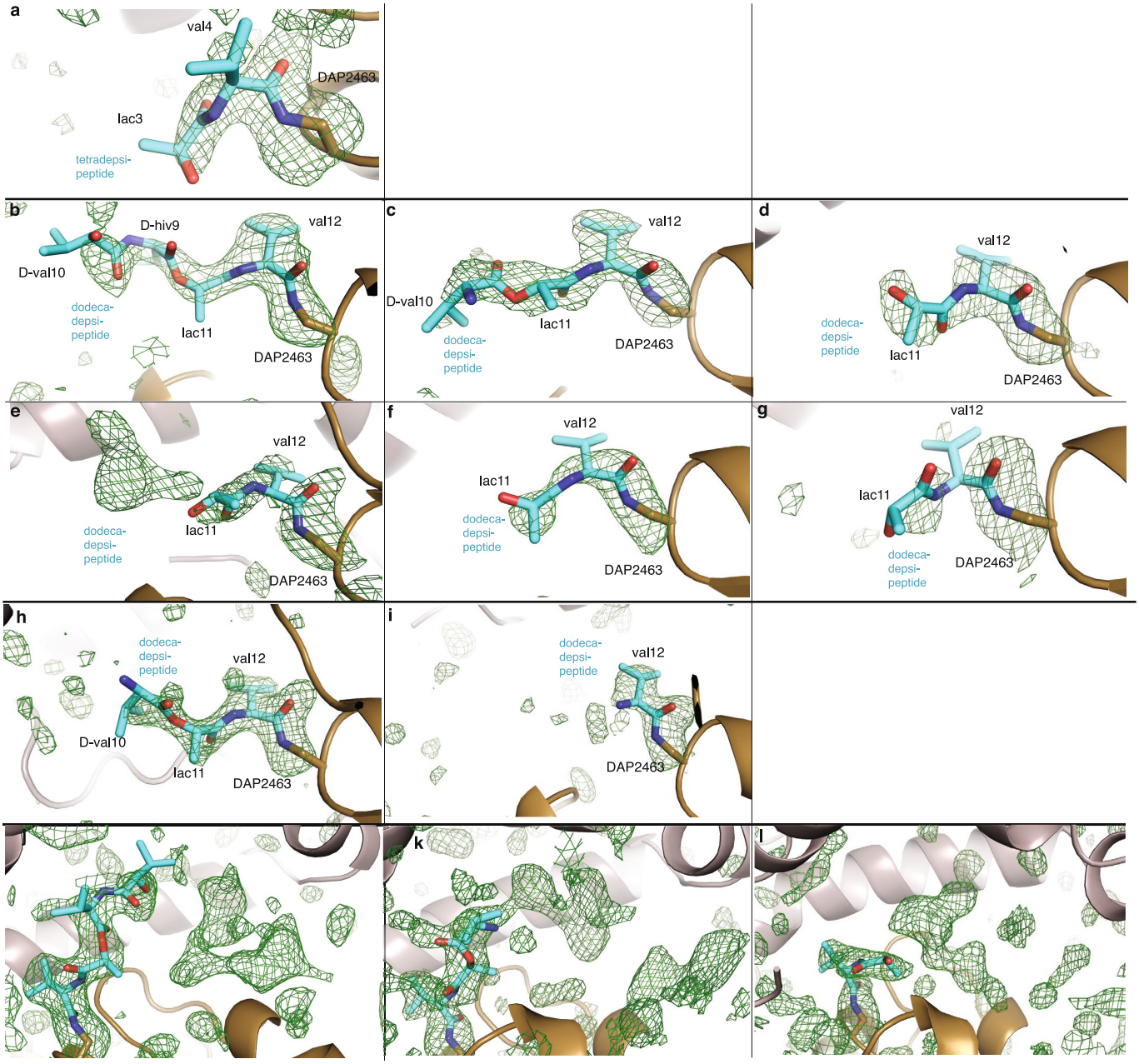
Extended Data Fig. 6 | Structures of Vlm TE_{wt} and tetradepsipeptidyl–TE_{DAP}, and top-down LC–ESI-MS of Vlm TE_{wt}. a, Secondary-structure elements of Vlm TE; the naming is based on the convention for α/β-hydrolase proteins. **b**, Comparison of two TE_{wt} structures (PDB accession numbers 6ECB and 6ECC). The active-site lid of the first structure (light grey) is nearly completely ordered, whereas the lid of second structure (dark grey) shows density for Lα3, Lα4 and Lα5 only. In the second structure, Lα3 is rotated 10° towards the active site. **c**, Deconvoluted mass spectra of TE_{wt} incubated with different substrates. Solid line, buffer control: expected molecular mass 31,028.22 Da; observed 31,028.75 Da. Dashed line, TE_{wt} incubated with tetradepsipeptidyl–SNAC: expected 31,028.22 Da (unmodified) and 31,399.44 Da (modified); observed 31,026.29 Da. Dotted line, TE_{wt} incubated with valinomycin: expected 31,028.22 Da (unmodified) and 32,139.86 Da (modified); observed 31,027.01 Da. Experiments were repeated independently twice with similar results. **d**, Comparison of near-identical conformations of TE_{wt} (light grey; 6ECB) and tetradepsipeptidyl–TE_{DAP} (tan and dark grey; 6ECD).

© 2019 Springer Nature Limited. All rights reserved.



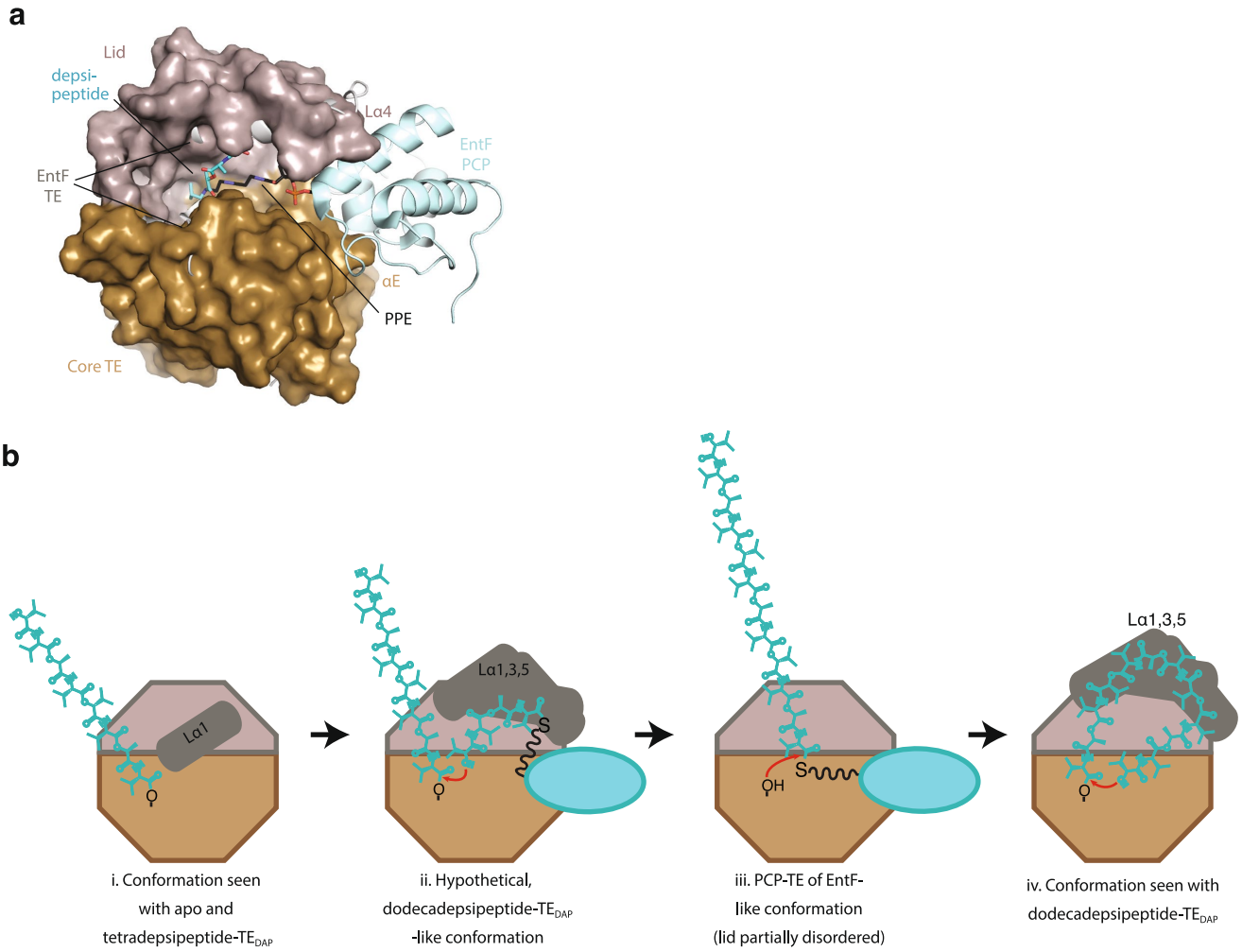
Extended Data Fig. 7 | Expression and substrate conjugation to Vlm TE containing DAP at position 2463. a, Following expression and purification of Vlm TE_{DAP}, the protein was loaded on an SDS–PAGE gel and Coomassie stained; the experiment was repeated in two biological replicates with similar results. **b**, The deprotection of **6** in TE_{DAP}–strep was followed by ESI-MS analysis. Green trace, purified TE_{DAP}–strep containing **6** at position 2463: expected mass 32,364.6 Da, observed 32,365.78 Da. Red trace, TE_{DAP}–strep containing **6** at position 2463 following illumination to convert **6** to the intermediate: expected 32,171.56 Da, observed 32,168.48 Da; and further incubation (1 h, 4 °C) to convert the intermediate to product: expected 32,067.62 Da, observed 32,068 Da). Blue trace, TE_{DAP}–strep containing **6** at position 2,463 following illumination (to convert **6** to the intermediate) and further incubation (10 h, 4 °C) to convert the intermediate to DAP (**1**): expected 32,067.62 Da; observed, 32,067.84 Da. The experiment was repeated in two biological replicates with similar results. **c**, Purified TE_{DAP} after illumination and intermediate fragmentation: expected 31,027.24 Da, observed 31,026.95 Da and 31,131.82 Da. **d**, TE_{DAP} incubated with tetradepsipeptidyl–SNAC **7**: expected 31,027.24 Da (unmodified) and 31,398.69 Da (modified); observed 31,025.92 Da and 31,396.55 Da. The experiments in **c**, **d** were repeated independently twice with similar results.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 8 | Electron density of the active site of covalent depsipeptidyl–TE_{DAP} complexes. Unbiased mF_o – DF_c maps (green mesh, contoured at 2.5σ), calculated before depsipeptide residues were placed in the model. DAP (brown) and depsipeptide residues (cyan) are depicted as sticks. **a**, Tetradepsipeptidyl–TE_{DAP} (PDB accession number 6ECD). **b**–**g**, Dodecadepsipeptidyl–TE_{DAP} P_1 space-group structure (6ECF), with crystallographically independent molecules A to F shown in sequential order. **h**, **i**, Dodecadepsipeptidyl–TE_{DAP} H_3 space group (6ECE), for crystallographically independent molecules A and B. **j**–**l**, Electron density of the active site of covalent depsipeptidyl–TE_{DAP} complexes extends beyond modelled depsipeptides. Unbiased mF_o – DF_c maps (green mesh, contoured at 2.5σ), calculated before depsipeptide residues were placed in the model, for dodecadepsipeptidyl–TE_{DAP} P_1 space-group structure, with crystallographically independent molecules A, B and D in sequential order. The observed electron density that extends beyond the modelled depsipeptides (cyan sticks) could accommodate extra depsipeptide residues in different orientations. However, unambiguous modelling into this density could not be achieved.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 9 | Modelling of interaction between the PCP domain and TE domain and putative pathway. a, Superimposition of dodecadepsipeptidyl–TE$_{DAP}$ with the structure of the EntF PCP–TE didomain[32] (PDB accession number 3TEJ) shows the path of the PPE moiety to the active site. **b**, Hypothetical pathway for oligomerization and cyclization, starting from octadepsipeptidyl–TE. i, The position of Lα1 in the observed apo/tetradepsipeptide conformation promotes an extended peptide conformation. ii, The tetradepsipeptidyl–PCP accepts the octadepsipeptide onto its terminal hydroxyl, perhaps using a dodecadepsipeptide-like lid conformation which could accommodate the roughly 30-Å tetradepsipeptidyl–PPE bound to the PCP domain and guide it towards the active site. iii, The PCP domain presents the thioester for transfer back to serine 2463. iv, Finally, the lid conformation observed in the dodecadepsipeptide–TE$_{DAP}$ structures could help to curl the dodecadepsipeptide back towards serine 2463 for cyclization.

32. Liu, Y., Zheng, T. & Bruner, S. D. Structural basis for phosphopantetheinyl carrier domain interactions in the terminal module of nonribosomal peptide synthetases. *Chem. Biol.* **18**, 1482–1488 (2011).

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Data collection and refinement statistics for the crystal structures presented here

	TE_{wt} structure 1 (**6ECB**)	TE_{wt} structure 2 (**6ECC**)	TE_{DAP} bound with a tetradepsipeptide (**6ECD**)	TE_{DAP} bound with dodecadepsipeptide, space group H3 (**6ECE**)	TE_{DAP} bound with dodecadepsipeptide, space group P1 (**6ECF**)
Data collection					
Space group	P 4 3 2	P 4 3 2	P 4 3 2	R 3 : H	P1
Cell dimensions					
a, *b*, *c* (Å)	151.4, 151.4, 151.4	152.2, 152.2, 152.2	152.3, 152.3, 152.3	77.6, 77.6, 235.2	77.0, 77.1,90.3
α, β, γ (°)	90, 90, 90	90, 90, 90	90, 90, 90	90, 90, 120	91.8, 114.9, 118.0
Resolution (Å)	151.4-1.7 (1.73-1.7)	152.2-1.8 (1.84-1.8)	107.7-1.9 (1.94-1.9)	64.59-2.0 (2.05-2.0)	78.55-2.5 (2.589-2.5)
R_{sym} or R_{merge}	0.063 (1.059)	0.08 (1.525)	0.123 (3.917)	0.079 (0.845)	0.071 (0.320)
I / sI	21.83 (1.3)	19.0 (1.1)	23.3 (1.0)	9.8 (1.5)	4.06 (1.85)
Completeness (%)	98.3 (88.3)	99.4 (94.2)	100 (100)	100 (100)	97.58 (97.10)
Redundancy	12.3 (4.5)	12.7 (6.2)	37.2 (37.4)	5.0 (4.9)	1.7 (1.7)
Refinement					
Resolution (Å)	75.7-1.7	87.87-1.8	107.7-1.9	44.24-2.0	78.55-2.5
No. reflections	64286	55853	48056	35677	53933
R_{work} / R_{free}	0.1725/0.1874	0.1757/0.1898	0.1879/0.2143	0.2091/0.2426	0.1969/0.2489
No. atoms	2188	1917	1918	3739	11761
Protein	1984	1746	1807	3645	11518
Ligand/ion	0	0	5	5	37
Water	204	171	106	89	206
B-factors	33.45	39.71	54.47	48.06	44.16
Protein	32.52	38.92	54.39	48.09	44.13
Ligand/ion	n/a	n/a	117.62	52.13	65.2
Water	42.47	47.77	52.94	46.47	41.81
R.m.s. deviations					
Bond lengths (Å)	0.009	0.019	0.018	0.005	0.005
Bond angles (°)	1.31	1.69	1.65	0.96	0.99

Each dataset was collected from a single crystal. Values in parentheses are for the highest-resolution shell.

© 2019 Springer Nature Limited. All rights reserved.

LETTER

https://doi.org/10.1038/s41586-018-0779-6

Structure of *Plasmodium falciparum* Rh5–CyRPA–Ripr invasion complex

Wilson Wong[1,2,6], Rick Huang[3,6], Sebastien Menant[1], Chuan Hong[3], Jarrod J. Sandow[1,2], Richard W. Birkinshaw[1,2], Julie Healer[1,2], Anthony N. Hodder[1,2], Usheer Kanjee[4], Christopher J. Tonkin[1,2], Denise Heckmann[1,2], Vladislav Soroka[5], Teit Max Moscote Søgaard[5], Thomas Jørgensen[5], Manoj T. Duraisingh[4], Peter E. Czabotar[1,2], Willem A. de Jongh[5], Wai-Hong Tham[1,2], Andrew I. Webb[1,2], Zhiheng Yu[3] & Alan F. Cowman[1,2]*

***Plasmodium falciparum* causes the severe form of malaria that has high levels of mortality in humans. Blood-stage merozoites of *P. falciparum* invade erythrocytes, and this requires interactions between multiple ligands from the parasite and receptors in hosts. These interactions include the binding of the Rh5–CyRPA–Ripr complex with the erythrocyte receptor basigin[1,2], which is an essential step for entry into human erythrocytes. Here we show that the Rh5–CyRPA–Ripr complex binds the erythrocyte cell line JK-1 significantly better than does Rh5 alone, and that this binding occurs through the insertion of Rh5 and Ripr into host membranes as a complex with high molecular weight. We report a cryo-electron microscopy structure of the Rh5–CyRPA–Ripr complex at subnanometre resolution, which reveals the organization of this essential invasion complex and the mode of interactions between members of the complex, and shows that CyRPA is a critical mediator of complex assembly. Our structure identifies blades 4–6 of the β-propeller of CyRPA as contact sites for Rh5 and Ripr. The limited contacts between Rh5–CyRPA and CyRPA–Ripr are consistent with the dissociation of Rh5 and Ripr from CyRPA for membrane insertion. A comparision of the crystal structure of Rh5–basigin with the cryo-electron microscopy structure of Rh5–CyRPA–Ripr suggests that Rh5 and Ripr are positioned parallel to the erythrocyte membrane before membrane insertion. This provides information on the function of this complex, and thereby provides insights into invasion by *P. falciparum*.**

The invasion of *P. falciparum* merozoites into erythrocytes requires the ligand Rh5, which binds to the host receptor basigin[1]. Rh5 forms a ternary complex with Ripr and CyRPA at the merozoite–erythrocyte interface[2,3]. This complex is linked to the formation of a pore between the merozoite and erythrocyte membrane, through which Ca^{2+} can pass[2,4].

To understand the function of the Rh5–CyRPA–Ripr complex, we expressed recombinant forms of Ripr, Rh5 and CyRPA proteins (Fig. 1a). The complex formed by Ripr, Rh5 and CyRPA migrates at 480 kDa in blue native electrophoresis, compared to Ripr alone, which has an apparent molecular weight of 242 kDa (Fig. 1a). We used anion-exchange chromatography to separate uncomplexed Ripr from ternary complexes; the peak contained the three proteins Rh5, CyRPA and Ripr, which confirms that a stable complex was formed (Fig. 1b). The recombinant Rh5–CyRPA–Ripr complex was equivalent in molecular weight to the endogenous complex purified from *P. falciparum*, as both migrated at 480 kDa (Fig. 1c). Chemically cross-linked Rh5–CyRPA–Ripr complex migrated at 212 kDa (Extended Data Fig. 1a), which indicates a 1:1:1 stoichiometric ratio: this ratio suggest that the migration on native PAGE was due to an elongated shape, which we confirmed using negative-stain electron microscopy (Extended Data Fig. 1b).

The basigin ectodomain that lacks the transmembrane region did not bind the Rh5–CyRPA–Ripr complex in solution, but did bind Rh5 (Fig. 1d, Extended Data Fig. 1c, d). However, full-length basigin that includes the lipid-embedded transmembrane region bound the ternary complex (Fig. 1d, Extended Data Fig. 1e, f). The affinity of interaction showed that Rh5, Rh5–CyRPA and the Rh5–CyRPA–Ripr complex all bound to full-length basigin with a similar affinity, of 200 nM (Fig. 1e, Extended Data Table 1). However, the Rh5–basigin and Rh5–CyRPA–Ripr–basigin interactions had an additional higher-affinity state of 50 nM that had a slower off-rate (Fig. 1e, Extended Data Table 1). This higher-affinity state was more prevalent in Rh5–CyRPA–Ripr–basigin than Rh5–basigin interactions, as reflected by the ratio of the low-affinity dissociation constant (K_{d1}) to the high-affinity dissociation constant (K_{d2}), which suggests that there were considerable conformational changes (Extended Data Table 2). These conformational changes in Rh5 were confirmed using hydrogen–deuterium exchange mass spectrometry, which showed that the disulfide loop (Cys345–Cys351) that forms part of the basigin-binding site[6] had a bimodal distribution of deuterium exchange that was consistent with two states; the detected protein sequence also underwent considerable conformational changes (Fig. 2a). Therefore, Rh5 undergoes conformational changes during binding to basigin that are stabilized in the ternary complex, and require interaction with lipid micelles surrounding the transmembrane helix of the receptor for efficient binding (Extended Data Fig. 1c–f).

We next showed the Rh5–CyRPA–Ripr complex bound to basigin on the erythroid cell line JK-1[7] (Fig. 2b, Extended Data Figs. 1h–i, 2). Rh5 bound to JK-1 cells in a basigin-dependent manner, with an approximately twofold-higher binding compared to JK-1 cells in which basigin has been deleted (JK-1ΔBSG cells) (Fig. 2b). Neither Ripr nor CyRPA bound to JK-1 or JK-1ΔBSG cells (Fig. 2b). Rh5–CyRPA did not show a significant level of binding to JK-1, which suggests that this assay detects high-affinity binding events and that in the binary complex CyRPA interferes with the interaction of Rh5 with basigin (Fig. 2b). Rh5–CyRPA–Ripr was detected on the surface of JK-1 cells at higher levels than was Rh5 alone, which indicates that the ternary complex bound JK-1 cells at significantly higher efficiency, consistent with the relative contribution of high-affinity binding sites (Fig. 1e, Extended Data Table 2). Additionally, the number of JK-1 cells detected that contain bound Ripr increased markedly for the ternary complex relative to Rh5 alone or the binary complex, which indicates that the association of Ripr with JK-1 cells was dependent on its presence in the complex. Although an increased level of CyRPA could be detected on JK-1 cells when bound in the ternary complex, this level was not as significant as for Rh5 and Ripr; this suggests that CyRPA dissociates from the complex during binding to basigin. Therefore, the binding of Rh5–CyRPA–Ripr to basigin initiates molecular events that mediate an increased association between Rh5–Ripr and erythrocyte membranes.

Owing to the requirement of lipid micelles for interactions between Rh5–CyRPA–Ripr and basigin (Fig. 1d, Extended Data Fig. 1e, f), we hypothesized that the ternary complex inserts into erythrocyte

[1]Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia. [2]Department of Medical Biology, University of Melbourne, Parkville, Victoria, Australia. [3]CryoEM Shared Resources, Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA, USA. [4]Department of Immunology and Infectious Diseases, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [5]ExpreS2ion Biotechnologies, Horsholm, Denmark. [6]These authors contributed equally: Wilson Wong, Rick Huang. *e-mail: cowman@wehi.edu.au

© 2019 Springer Nature Limited. All rights reserved.

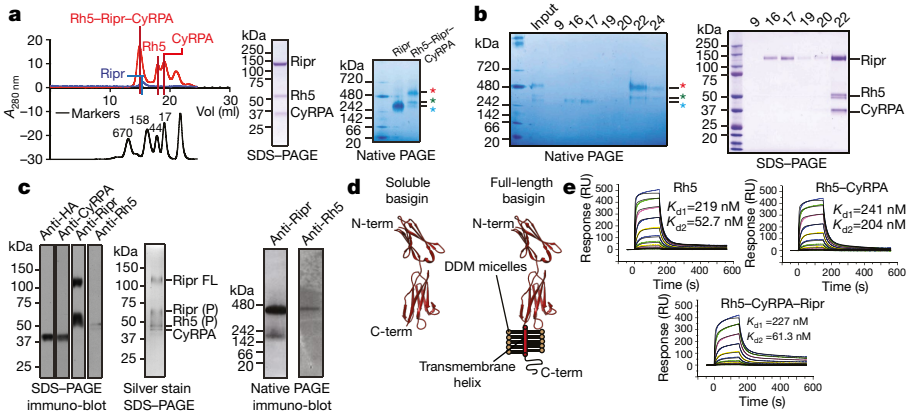


Fig. 1 | Rh5, CyRPA and Ripr form a ternary complex. a, Size-exclusion chromatography peaks of Rh5–CyRPA–Ripr (red), or Ripr (blue). Red asterisk, Rh5–CyRPA–Ripr; green asterisk, CyRPA–Ripr; blue asterisk, Ripr. SDS–PAGE, sodium dodecyl sulfate polyacrylamide gel electrophoresis. **b**, Anion-exchange chromatography elution of Rh5–CyRPA–Ripr, and the separation of Ripr. **c**, Purification of Rh5–CyRPA–Ripr from parasites. (P) denotes bands from protease processing of Ripr. FL, full length. Experiments in **a**–**c** were repeated at least three times with biologically independent samples, and were reproducible. For gel and western blot source data, see Supplementary Fig. 1a, b and c (which correspond to **a**, **b** and **c**, respectively). **d**, Ectodomain (left) and full-length basigin (right) with transmembrane and *n*-dodecyl-β-D-maltoside (DDM) micelle. C-term, C terminus; N-term, N terminus. **e**, Surface plasmon resonance measuring the interaction of Rh5, Rh5–CyRPA and Rh5–CyRPA–Ripr complexes with full-length basigin. RU, resonance unit.

membranes upon binding to basigin. To test this, we measured the ability of proteins to lyse erythrocytes as an indication of membrane insertion activity[8]. Lysis activity was observed when erythrocytes were incubated with the Rh5–CyRPA–Ripr complex (Fig. 2c), whereas no significant activity was detected with single or binary components. Therefore, Rh5–CyRPA–Ripr disrupts the erythrocyte membrane using excess, non-physiological concentration of proteins; however, the concentration of the ternary complex during merozoite invasion would be precisely controlled, and not result in erythrocyte lysis.

Differential solubility in detergent was used to confirm the insertion of proteins into erythrocyte membranes (Fig. 2d). For Rh5, Rh5–CyRPA and ternary complexes, a proportion of the Rh5 pool was in the Triton-X100 detergent-resistant membrane fraction (Fig. 2d, Extended Data Fig. 1j), which indicates a high-molecular-weight species embedded within a highly ordered and condensed membrane. When added as a ternary complex, Ripr was also found in detergent-resistant membrane fractions, which indicates it also was inserted into the membrane. However, CyRPA was in the soluble fraction. This suggests that after binding to basigin, the Rh5–CyRPA–Ripr complex disassembles; CyRPA is excluded from the membrane, whereas Rh5 and Ripr are inserted into the membrane (Fig. 2d). Rh5–Ripr associated with the erythrocyte membrane migrated as a single band of high-molecular weight (about 700 kDa), indicating that they were in the same complex as the oligomers (Fig. 2e). However, the insertion of Rh5–Ripr into the membrane did not alter its permeability to Ca^{2+} (Extended Data Fig. 3), which suggests that additional proteins are required for pore formation between erythrocyte membranes and the apical end of invading merozoites[2,4].

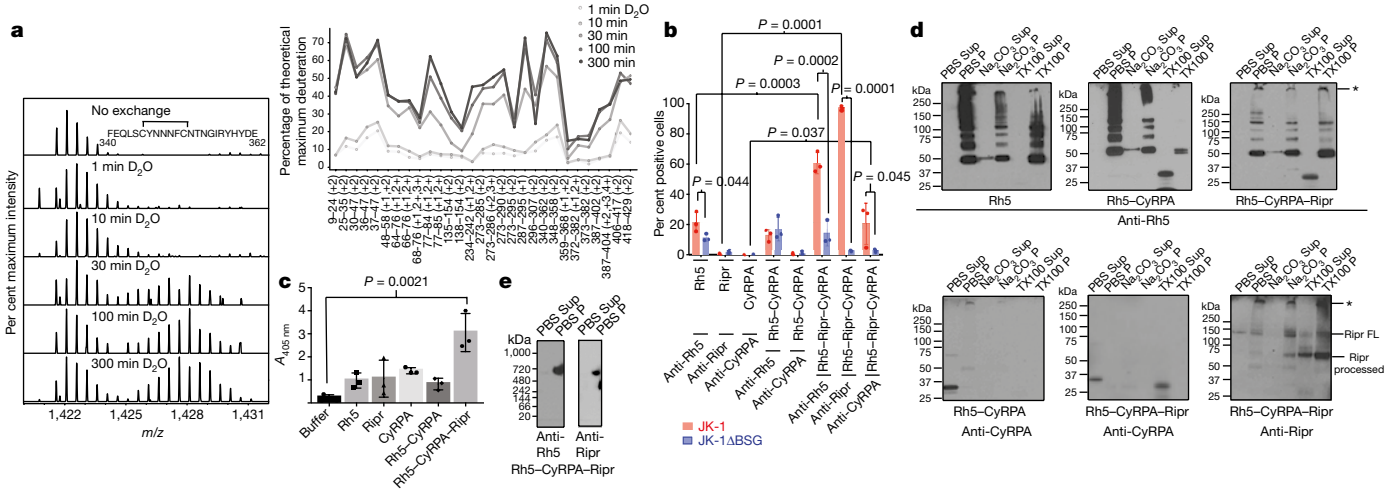


Fig. 2 | The Rh5–CyRPA–Ripr complex inserts into membranes. **a**, Hydrogen–deuterium exchange mass spectrometry analysis of Rh5, showing deuterium incorporation across the peptide-spanning disulfide loop Cys345–Cys351 (left) and detected peptides (right). **b**, Fluorescence-activated cell sorting (FACS) analyses of Rh5, Ripr, CyRPA, Rh5–CyRPA and Rh5–CyRPA–Ripr binding to JK-1 and JK-1ΔBSG cells. A plot of the percentage of positive cells detected after incubation with protein(s) is shown. **c**, Haemolytic activity of Rh5, CyRPA, Ripr, Rh5–CyRPA and Rh5–CyRPA–Ripr complexes. For **b**, **c**, $n = 3$; experiments were performed at least 3 times with biologically independent samples and were reproducible. Bar graphs show mean values with standard deviation. Student's *t*-test was used to calculate statistical significance with two-tailed *P* value. **d**, Differential solubilization of Rh5, Rh5–CyRPA and Rh5–CyRPA–Ripr with erythrocytes. Samples were separated on non-reducing SDS–PAGE, analysed by western blot. Asterisk denotes high-molecular weight species. PBS, Phosphate buffered saline. **e**, Pelleted erythrocyte membranes after an insertion of Rh5 and Ripr, detected by native-PAGE immuno-blotting analyses. For **d**, **e**, experiments were repeated 3 times with biologically independent samples, and were reproducible. For western blot source data, see Supplementary Fig. 2d, e.

© 2019 Springer Nature Limited. All rights reserved.

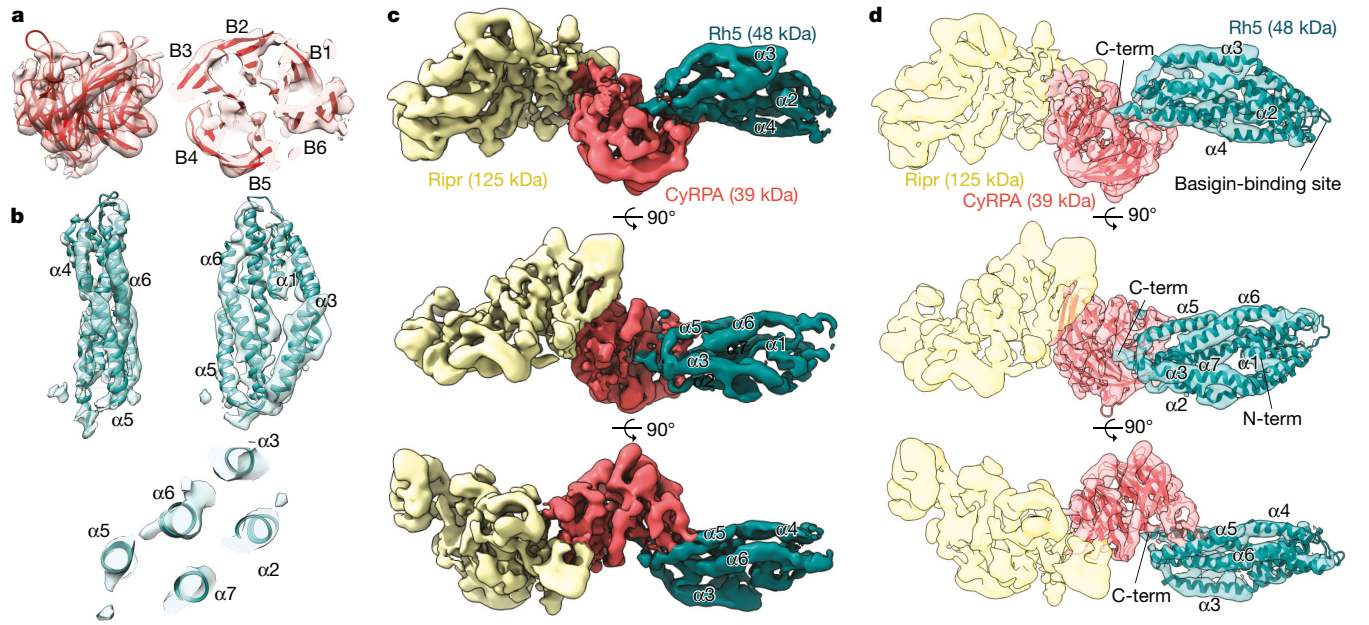


Fig. 3 | Organization of Rh5–CyRPA–Ripr ternary complex. a, Electron microscopy density of the CyRPA region of the Rh5–CyRPA–Ripr complex (left) and a cross-section that shows the resolution of 6-bladed β-sheets (right). **b**, Electron microscopy density of the Rh5 region of the Rh5–CyRPA–Ripr complex (top), and a cross-section that shows the resolution of 5 α-helices (bottom). **c**, **d**, 3D reconstruction of Rh5–CyRPA–Ripr at global resolution of 7.17 Å, with map shown in **c** and refined model in **d**.

We used cryo-electron microscopy (cryo-EM) to obtain structural insights into the Rh5–CyRPA–Ripr complex (Extended Data Fig. 4). Three-dimensional classification resulted in the separation of two populations, corresponding to the CyRPA–Ripr and Rh5–CyRPA–Ripr complexes (Extended Data Fig. 4c). Fourier shell correlation reported the global resolution of binary and ternary complexes at 5.07 Å and 7.17 Å, respectively (Extended Data Fig. 4d, e). Local resolution suggested that Rh5 was flexible (with the basigin-binding site being the most flexible), whereas the CyRPA and Ripr regions were more stable (Extended Data Fig. 4f, g, Extended Data Table 3). Densities corresponding to the six-blade β-sheets of the CyRPA β-propeller in the binary map were resolved, including several β-strands within blades 1, 3 and 6 of the β-propeller (Extended Data Fig. 5a, b). In the map of the ternary complex, densities that correspond to the six-blade β-sheets of the CyRPA β-propeller[9,10] were resolved, as were the six α-helices of Rh5[5,6] (Fig. 3a, b, Extended Data Fig. 5c).

The Rh5–CyRPA–Ripr complex was composed of a stoichiometric ratio of 1:1:1 with an elongated shape, in which CyRPA constitutes the core that stabilizes Rh5 and Ripr on the opposite sides of the ternary complex (Fig. 3c, d, Extended Data Fig. 1a, Supplementary Video 1). The basigin-binding site of Rh5—consisting of the α2 and α4 helices, and the disulfide loop (Cys345–Cys351)[6]—was located at the tip of the ternary complex opposite to Ripr, which is solvent-exposed; thus, CyRPA and Ripr did not contact basigin (Fig. 3c, d). This was consistent with the K_d values of Rh5 and Rh5–CyRPA–Ripr complex for basigin being similar, which demonstrates that the ternary complex interacts with basigin via Rh5 (Fig. 1e).

The CyRPA-binding site of Rh5 was at the tip of the α-helical scaffold, opposite the basigin-binding site (Fig. 4a). Density for the C-terminal tail of Rh5 and part of the α7 helix was inserted into the central cavity of the CyRPA β-propeller (Fig. 4a, Extended Data Fig. 5d). At this contact site, the α5 and α7 helices of Rh5 present to CyRPA a hydrophobic groove enriched in hydrophobic residues (Fig. 4a, Extended Data Fig. 5e, f). Two loops (the B4 loop and B4–B5 connecting loop) presented by blades 4 and 5 of the CyRPA β-propeller, which are enriched with several aromatic residues (Tyr185, Phe187 and Phe226), are inserted in this groove of Rh5 (Fig. 4a, Extended Data Fig. 5e, f). Upon the binding of Rh5 to CyRPA, it is likely the B5 loop located on blade 5 of CyRPA becomes disordered to accommodate occupancy by the α5 helix of Rh5 (Extended Data Fig. 5g). Cross-linking studies detected an interaction between the C terminus of Rh5 that immediately precedes helix α7 with blade 1 of CyRPA, consistent with our cryo-EM structure (Fig. 4, Extended Data Fig. 5h).

Although the Ripr model could not be built de novo owing to resolution of the map and the presence of primarily β-sheet structures (Extended Data Fig. 6a), several secondary structural elements were clearly visible at the contact interface of CyRPA–Ripr within the ternary complex. Secondary structure predictions indicated two putative α-helices (residues 196–211 and residues 364–373) residing in the N terminus of Ripr (Extended Data Fig. 6b) and the rest of the amino acid sequence at the C terminus (residues 373–1086) was predicted to contain loops and β-strands, including eight epidermal growth factor-like repeats (EGFs 3–10). At the interface of CyRPA–Ripr, density for a four-turn α-helix that contacts blade 6 of the CyRPA β-propeller could be observed (Fig. 4b). The length of this α-helix density suggests it corresponds to residues 196–211 of Ripr. In addition, density for a β-strand of Ripr forms an intermolecular β-sheet interaction with blade 6 of the CyRPA β-propeller (Fig. 4b). Therefore, blades 4 and 5 of the CyRPA β-propeller provide contact sites for Rh5, and blade 6 provides a contact site for Ripr (Fig. 4c).

A crystal structure of Rh5–basigin complex has previously been published[6], which enabled alignment of the Rh5–CyRPA–Ripr cryo-EM structure to the Rh5–basigin crystal structure (Extended Data Fig. 6c). The superimposed structures suggest that the Rh5–CyRPA–Ripr complex was positioned parallel to the erythrocyte membrane (Fig. 4d). This orientation in relation to the erythrocyte membrane—along with the conformationl changes detected when the ternary complex is bound to basigin (Fig. 1e)—could facilitate the membrane insertion of Rh5 and Ripr. The C-terminal helical bundle of Rh5 is structurally similar to the N-terminal coiled-coil domain of SipB (root mean square deviation of 3.4 Å over 144 residues) of the bacterial type III secretion system, and possesses an amphipathic property[11] (Extended Data Fig. 6d, e). The membrane-inserted Rh5–Ripr complex, along with other as-yet unidentified parasite proteins, may be involved in the formation of a pore that enables invading merozoites to inject components into the erythrocyte cytoplasm.

© 2019 Springer Nature Limited. All rights reserved.

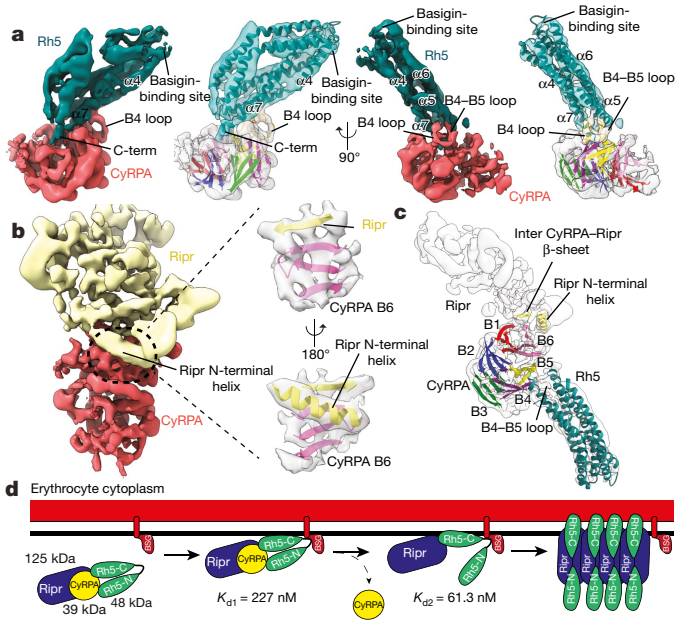


Fig. 4 | Interactions between Rh5–CyRPA and CyRPA–Ripr. a, Electron microscopy density of CyRPA (red) bound to Rh5 (blue). B4 and B4–B5 loops of CyRPA were inserted into the hydrophobic groove of Rh5 formed by α5 and α7 helices. **b**, Electron microscopy density of CyRPA (red) bound to Ripr (yellow) (left). Contact between CyRPA and Ripr magnified (right), showing the intermolecular β-sheet interaction and binding of the Ripr N-terminal α-helix to blade 6 of CyRPA β-propeller. **c**, Model of Rh5–CyRPA–Ripr showing contacts between Rh5–CyRPA and CyRPA–Ripr. **d**, Model of molecular events for binding and insertion of Rh5–CyRPA–Ripr complex. The ternary complex binds basigin via a lower-affinity binding site, in an interaction that requires membrane lipid. Upon initial interaction with basigin, a conformational change in Rh5 leads to a high-affinity interaction and exposure of an amphipathic helical domain in Rh5. This leads to oligomerization and insertion of Rh5 and Ripr into the erythrocyte membrane, whereas CyRPA is excluded.

It is likely that the conformational changes observed for Rh5 in the ternary complex can be blocked during merozoite invasion by inhibitory antibodies. A monoclonal antibody to Rh5 (9AD4) has previously been identified[12]. This monoclonal antibody does not block the basigin–Rh5 interaction but does inhibit invasion; it may act by interference, producing conformational changes that block the function of the complex (Extended Data Fig. 6c). Additionally, other monoclonal antibodies that bind Rh5 and CyRPA could sterically interfere with the docking of the Rh5–CyRA–Ripr complex to the erythrocyte membrane, preventing the membrane insertion of Rh5 and Ripr[5,10] (Extended Data Fig. 6c). This raises a route through which to identify epitopes for antibodies that block conformational changes and membrane insertion, which would trap the complex in an inactive state and inhibit invasion (Supplementary Discussion). Collectively, these data lay the foundation for understanding the molecular details of the invasion of *P. falciparum* into erythrocytes, which will be important for designing vaccines against this disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0779-6.

Received: 23 August 2017; Accepted: 25 October 2018;
Published online 12 December 2018.

1. Crosnier, C. et al. Basigin is a receptor essential for erythrocyte invasion by *Plasmodium falciparum*. *Nature* **480**, 534–537 (2011).
2. Volz, J. C. et al. Essential role of the PfRh5/PfRipr/CyRPA complex during *Plasmodium falciparum* invasion of erythrocytes. *Cell Host Microbe* **20**, 60–71 (2016).
3. Chen, L. et al. An EGF-like protein forms a complex with PfRh5 and is required for invasion of human erythrocytes by *Plasmodium falciparum*. *PLoS Pathog.* **7**, e1002199 (2011).
4. Weiss, G. E. et al. Revealing the sequence and resulting cellular morphology of receptor–ligand interactions during *Plasmodium falciparum* invasion of erythrocytes. *PLoS Pathog.* **11**, e1004670 (2015).
5. Chen, L. et al. Crystal structure of PfRh5, an essential *P. falciparum* ligand for invasion of human erythrocytes. *eLife* **3**, (2014).
6. Wright, K. E. et al. Structure of malaria invasion protein RH5 with erythrocyte basigin and blocking antibodies. *Nature* **515**, 427–430 (2014).
7. Kanjee, U. et al. CRISPR/Cas9 knockouts reveal genetic interaction between strain-transcendent erythrocyte determinants of *Plasmodium falciparum* invasion. *Proc. Natl Acad. Sci. USA* **114**, E9356–E9365 (2017).
8. Blocker, A. et al. The tripartite type III secreton of *Shigella flexneri* inserts IpaB and IpaC into host membranes. *J. Cell Biol.* **147**, 683–693 (1999).
9. Chen, L. et al. Structural basis for inhibition of erythrocyte invasion by antibodies to *Plasmodium falciparum* protein CyRPA. *eLife* **6**, e21347 (2017).
10. Favuzza, P. et al. Structure of the malaria vaccine candidate antigen CyRPA and its complex with a parasite invasion inhibitory antibody. *eLife* **6**, e20383 (2017).
11. Barta, M. L. et al. The structures of coiled-coil domains from type III secretion system translocators reveal homology to pore-forming toxins. *J. Mol. Biol.* **417**, 395–405 (2012).
12. Douglas, A. D. et al. Neutralization of *Plasmodium falciparum* merozoites by antibodies against PfRH5. *J. Immunol.* **192**, 245–258 (2014).

Acknowledgements We thank L. Chen, J. Thompson, E. Hanssen, A. Leis and P. de Fonseca for experimental assistance, and the Victorian Red Cross Blood Bank for blood. The data presented here were made possible through Victorian State Government Operational Infrastructure Support and Australian Government NHMRC IRIISS. The research was directly supported by a National Health and Medical Research Council of Australia (NHMRC).

Reviewer information *Nature* thanks L. Miller, S. Scheres, A. Sharma and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions W.W. performed biochemistry and cryo-EM, built the atomic model and wrote the manuscript. R.H., C.H. and Z.Y. performed cryo-EM and analysis. J.H., A.N.H., V.S., T.M.M.S., T.J. and W.A.d.J. performed protein purification. R.W.B. and P.E.C. performed plasmon surface resonance. S.M. and W.-H.T. performed FACS analysis. C.J.T. performed Ca^{2+} uptake experiments. D.H., J.J.S. and A.I.W. performed mass spectrometry. U.K. and M.T.D. made JK-1ΔBSG and JK-1 cells. All authors assisted with manuscript preparation. A.F.C. was responsible for project strategy, management, data interpretation and writing the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0779-6.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0779-6.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to A.F.C.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© 2019 Springer Nature Limited. All rights reserved.

METHODS

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Protein expression and purification. The expression and purification of recombinant Rh5 and CyRPA have previously been described[5,9]. Full-length *P. falciparum* Ripr (amino acids 20–1086) was expressed in *Drosophila* S2 cells (ExpreS^{2}ion Biotechnologies) and purified using the Strep-TactinXT purification system[13]. Full-length basigin was expressed in SF21 cells as per the supplier's manual, as a C-terminally Flag-tagged fusion protein. Full-length basigin was extracted from the membrane of SF21 cells in lysis buffer (40 mM Tris, 150 mM NaCl, 1% DDM, pH 8.5) and clarified supernatant, containing DDM-solublized basigin, was incubated with Flag resins at 4 °C for 2 h to enable binding. Basigin-bound resins were washed and eluted in elution buffer (20 mM Tris, 150 mM NaCl, 0.4 mM DDM and 100 μg/ml Flag peptides, pH 8.5). Eluted fractions contatining full-length basigin were further purified by size-exclusion chromatography using a Superose 6 10/300 size-exclusion column in elution buffer (20 mM Tris, 150 mM NaCl, 0.4 mM DDM, pH 8.5).

For preparation of the Rh5–CyRPA–Ripr complex, individual components were mixed at 1:1:1 molar ratio at room temperature for 1 h. The sample was injected onto a Superose 6 10/300 size-exclusion column in elution buffer (20 mM Tris, 250 mM NaCl, pH 8.5), and the Rh5–CyRPA–Ripr complex was separated from uncomplexed components. Fractions that contained the eluted Rh5–CyRPA–Ripr complex were pooled and diluted in ion-exchange buffer A (20 mM Tris, 25 mM NaCl, pH 8.5) followed by loading of the sample onto a HiTrap Q HP column. After an extensive wash in buffer A, the Rh5–CyRPA–Ripr complex was eluted in a linear gradient of buffer A and buffer B (20 mM Tris, 1 M NaCl, pH 8.5). This resulted in the separation of free Ripr from the ternary complex.

Surface plasmon resonance. Surface plasmon resonance binding assays were performed using BIAcore 4000 instruments in an SPR buffer (10 mM HEPES, 150 mM NaCl, 3.4 mM EDTA, 0.005% Tween-20, pH 7.4). Basigin—including the transmembrane helix—was immobilized as the ligand on a CM5 sensor chip surface by amine coupling. Hydrodynamic addressing was used to immobilize basigin on spots 1 and 5 of a single flow cell at densities of 7,725 RU and 3,834 RU, respectively. Sensograms were double-referenced by subtracting spot 2 or spot 4, which had been blocked with ethanolamine (spot 2 from 1, or spot 4 from 5), and a blank SPR buffer-only sample. Analyte protein samples (Rh5 alone, Rh5–CyRPA binary and Rh5–CyRPA–Ripr ternary complexes) were reconstituted in SPR buffer at various concentrations (2 μM to 2 nM) to derive binding affinities and injected for 150 s, with a 500-s dissociation time. The sensor surface was regenerated with glycine buffer (10 mM glycine pH 2.1) between each cycle before repeating analyte injections. Sensorgrams were initially fitted to a Langmuir specific one-site binding model and then a heterogeneous ligand-binding model, if appropriate, to derive on- and off-rates and the K_d values.

To determine the affinity of basigin binding to the Rh5-only, Rh5–CyRPA binary and Rh5–CyRPA–Ripr ternary complexes, SPR experiments were performed, immobilizing basigin on the sensor surface and flowing the various Rh5 complexes as the analyte. Because Rh5 and Ripr do not interact in the absence of CyRPA (Extended Data Fig. 1g), the binding affinity of the Rh5–Ripr binary complex for basigin could not be measured. Initially, the SPR curves were analysed by a 1:1 binding model, and then by a heterogeneous ligand model for the interaction. The 1:1 binding model gave K_d values of 240 nM, 180 nM and 130 nM for Rh5-only, Rh5–CyRPA and Rh5–CyRPA–Ripr, respectively (Extended Data Table 1). Visual inspection of the 1:1 binding model fit to the raw data indicated that Rh5-only and Rh5–CyRPA binary were in reasonable agreement with the data; however, the Rh5–CyRPA–Ripr ternary complex showed a poor fit. Hydrogen–deuterium exchange mass spectrometry (Fig. 2a) experiments provided evidence for two distinct populations of apo-Rh5 conformers, and electron-microscopy local-resolution analysis also indicated that the basigin-binding site is the most flexibile part of the ternary complex (Extended Data Fig. 4f). This provided justification for analysis by a heterogeneous ligand model, fitting the data to two independent sites with two distinct on- and off-rates (k_{on1}, k_{on2}, k_{off1} and k_{off2}) and providing one K_d value for each site (K_{d1} and K_{d2}). The Rh5–CyRPA–basigin binding showed minor differences between the two sites with similar K_d values, 200 and 240 nM (Fig. 1e, Extended Data Table 2), which were comparable to the values from the 1:1 binding model. However, both the Rh5–basigin and Rh5–CyRPA–Ripr–basigin interactions showed clear differences between the two sites with a low-affinity site that was comparable to the 1:1 binding model affinity (around 200 nM), and an additional higher-affinity binding site with slower on- and off-rates and a fourfold-lower K_d, of around 50 nM (Fig. 1e, Extended Data Table 2). This suggested two discrete conformations for Rh5 and the Rh5–CyRPA–Ripr ternary complex (consistent with the hydrogen/deuterium exchange mass spectrometry and electron microscopy data) that have differing affinities for basigin. Comparing the contribution of R_{max} to each fit showed that with Rh5 only, the low-affinity site dominated and the high-affinity site contributed 10% (9:1 ratio) of the fit (Extended Data Table 2). This ratio was decreased with the Rh5–CyRPA–Ripr ternary complex, with the high-affinity site contributing 30% (7:3 ratio) of the fit (Extended Data Table 2); this provides an explanation for the poor fit to the 1:1 binding model. Additionally, this indicates that the high-affinity conformation is stabilized in the ternary state.

Cell lines. *P. falciparum* strain 3D7 was obtained from D. Walliker at Edinburgh University, and was validated using whole-genome sequencing. The erythroleukaemia cell line JK-1 was obtained from the Leibniz Institute Deutsche Sammlung von Mikroorganismen und Zellkulturen collection of microorganisms and cell cultures (catalogue no. ACC347). The identity of the JK-1 cell line was confirmed by detection of specific proteins, such as basigin, on the surface. All cell lines tested negative for mycoplasma contamination.

Antibodies. Antibodes raised against Rh5, CyRPA and Ripr were generated in the Cowman laboratory and have previously been published[3,5,9].

Cell-binding assay based on flow cytometry. All binding and antibody incubations were performed at room temperature for 1 h in 50 μl. Washes were performed in phosphate-buffer saline (PBS) supplemented with 1% (w/v) bovine serum albumin (BSA) and spun at 1,000*g* for 1 min. In all conditions, an equivalent molarity of Rh5 at 4 μM was used. The binary Rh5–CyRPA and ternary Rh5–CyRPA–Ripr protein complexes were created by combining equimolar amounts of proteins in PBS and 1% BSA for one hour at room temperature. JK-1 and JK-1ΔBSG cells were washed twice and 1×10^6 cells per binding condition were used. Rh5, the binary Rh5–CyRPA and ternary Rh5–CyRPA–Ripr protein complexes were added separately to cells for 1 h at room temperature. After binding, the cells were washed and incubated with specific primary antibodies (0.2 mg/ml of monoclonal anti-Rh5, 0.05 mg/ml of polyclonal anti-Ripr or 0.05 mg/ml of polyclonal anti-CyRPA). After two washes, Alexa Fluor 488-conjugated goat anti-mouse or goat anti-rabbit secondary antibodies (1:100; Life Technologies) were added. The cells were washed three times and resuspended in 150 μl PBS before analysis with the LSRII flow cytometer (BD Biosciences). Thirty thousand events were recorded and results were analysed using the FlowJo software. The background signal induced by the primary and secondary antibodies in the absence of protein was subtracted from the corresponding positive fluorescent signals.

Haemolytic assay. Erythrocytes were washed in PBS 4 times before incubation with buffer control or 1.6 μM of purified recombinant proteins (Rh5, Ripr, CyRPA, Rh5–CyRPA binary and Rh5–CyRPA–Ripr ternary complexes) at 37 °C for 24 h with shaking. Unlysed cells were pelleted by centrifugation at 6,000 rpm for 1 min. The absorbance on the supernatant containing the released haemoglobin was measured at 405 nm.

Differential solubilization of proteins in erythrocyte membrane. Remaining cell pellets from the haemolytic assay were washed in PBS 4 times to release soluble proteins and subsequently pelleted by centrifugation at 6,000 rpm at 4 °C for 1 min. The PBS-washed cell pellets were treated with Na_2CO_3 pH 11.5 to release peripheral membrane associated proteins or Triton X100 to release integral membrane proteins. Centrifugation at 40,000 rpm at 4 °C for 20 min was performed to isolate the Na_2CO_3 and Triton X100 soluble and insoluble fractions for western blot analysis.

Ca^{2+} flux measurements using FACS. Erythrocytes were resuspended in Ringers buffer at 1% haematocrit and labelled with 5 μM of Fluo-4AM for 30 min at room temperature. Erythrocytes were then further diluted in Ringers buffer to 0.1% haematocrit and aliquoted to 200 μl in FACS tubes. Ca^{2+} ionophore A23187 was diluted in Ringers buffer to 2× final concentration. Rh5 alone or Rh5–CyRPA–Ripr were mixed and diluted in PBS. Fluorescence of Fluo-4-loaded erythrocytes were then acquired on a LSRII FACS analyser (BD Biosciences). Samples were analysed in FlowJo v.8.

Hydrogen–deuterium exchange mass spectrometry. Sample stock solutions were diluted to 40 pmol/μl protein concentration with 100 mM NaCl, 20 mM HEPES, pH 7.5. Two microlitres protein solution was transferred into a 10-mm autosampler vial (Thermo Scientific), with 38 μl of deuterium buffer and 2 μl of quench buffer (1.5% v/v formic acid) where these were used. Twelve microlitres acidified protein was injected into the sample loop and subsequently digested, desalted and separated online using the Agilent Technologies 1200 series Capillary LC System. The injected sample was delivered to an immobilized pepsin column (Poroszyme Immobilized Pepsin Cartridge, 2.1 mm × 30 mm, cat. number 2-3131-00, Applied Biosystems) at a flow rate of 50 μl/min buffer A1 (5% v/v methanol, 0.2% v/v formic acid in MiliQ water, pH 2.5) using an Agilent Technologies 1200 series pump, which equated to a digestion time of two minutes. The online digestion and subsequent separation steps were performed at 1 °C by storing lines, pepsin column, C18-trap and valve (Agilent Technologies 1200 series) in a 120-l fridge (Westinghouse). The flow was diverted by a two-position, ten-port valve and a binary pump (Agilent Technology 1200 series). The resulting peptic peptides were trapped on a C18 trap column (0.5 mm × 5 mm, ReproSil-Pur C18-AQ 5 μm, Dr. Maisch) and desalted with 95% buffer A2 (0.2% v/v formic acid in MiliQ water)

© 2019 Springer Nature Limited. All rights reserved.

and 5% buffer B2 (95% v/v acetonitrile, 0.2% v/v formic acid) at a flow rate of 5 μl/min. A 10-min linear gradient (5–55% buffer B2) starting after 3.2 min was applied to elute the peptides. The eluate was directed into a Thermo LTQ XL Hybrid Ion Trap-Orbitrap mass spectrometer with an ESI source operated at a capillary temperature of 180 °C, and a spray voltage of 1.8 kV using a 3-μm ID-conductive-coated pulled ESI emitter tip (New Objective). Mass spectra were acquired over the *m*/*z* range 350–2,000 using the Orbitrap analyser. For peptide identification, the five most-abundant ions per scan were fragmented and analysed in the ion trap. For each sample run, spectra were acquired for 20 min and the system was flushed and re-equilibrated after every sample measurement by injecting MiliQ water and performing a blank run. Each sample was analysed in triplicate. To identify peptides and determine sequence coverage, the acquired tandem mass spectrometry data were subjected to a protein database search, including a customized database that featured the sequence of Rh5 recombinant protein.

The hydrogen–deuterium exchange mass spectrometry analysis revealed that Rh5 exists in multiple different conformations. Peptides spanning the disulfide loop (Cys345–Cys351), which form part of the basigin-binding site[6], showed a bimodal distribution of deuterium exchange, which suggests that this region of the protein exists in two distinct conformational states (Fig. 2a). Additionally, most of the detected protein sequence appeared to undergo considerable exchange of deuterium over the time course, which suggests dynamic changes in most regions of the protein and an absence of disordered regions (Fig. 2a).

Cross-linking mass spectrometry. Protein samples were manually excised from preparative SDS–PAGE gels and subjected to manual in-gel reduction, alkylation and tryptic digestion. All gel samples were reduced with 10 mM DTT (Sigma) for 30 min, alkylated for 30 min with 50 mM iodoacetamide (Sigma) and digested with 375 ng trypsin gold (Promega) for 16 h at 37 °C. The extracted peptide solutions were then acidified (0.1% formic acid) and concentrated to 10 μl by centrifugal lyophilization using a SpeedVac AES 1010 (Savant). Extracted peptides were injected and fractionated by reversed-phase liquid chromatography on a nanoACQUITY UHPLC system (Waters) using a nanoACQUITY C18 250 mm × 0.075 mm I.D. column (Waters) with a linear 90-min gradient at a flow rate of 300 nl/min from 98% solvent A (0.1% formic acid in Milli-Q water) to 35% solvent B (0.1% formic acid, 99.9% acetonitrile). The nano-UHPLC was coupled online to a Q-Exactive Orbitrap mass spectrometer equipped with a nano-electron spray ionization source (Thermo Fisher Scientific). High mass-accuracy mass spectrometry data were obtained in a data-dependent acquisition mode with the Orbitrap resolution set at 70,000 and the top-ten multiply charged species selected for fragmentation by higher-energy collisional dissociation. The stepped (N)CE voltage was set to 19.5, 26 and 32.

Raw files were analysed using MaxQuant1,2 (version 1.5.3.30)[14,15]. The database search was performed using the Uniprot *P. falciparum* (isolate 3D7) database plus common contaminants, with strict trypsin specificity allowing up to two missed cleavages. The minimum peptide length was seven amino acids. Carbamidomethylation of cysteine was a fixed modification, and N-acetylation of the N termini of proteins and oxidation of methionine were set as variable modifications. During the MaxQuant main search, precursor ion mass error tolerance was set to 4.5 ppm and fragment ions were allowed a mass deviation of 20 ppm[14,15]. Peptide spectrum matches and protein identifications were filtered using a target–decoy approach at a false-discovery rate of 1%. MaxQuant APL files were converted to MGF files using the APL to MGF convertor software (https://www.wehi.edu.au/people/andrew-webb/1298/apl-mgf-converter)[14]. Cross-linked peptides were identified from the MGF files using StavroX software (version 3.6.0.1)[15]. Lysines, protein N termini, serines, threonines and tyrosines were set as reaction sites of the cross-linker NHS–esters. Trypsin was set as the enzyme, allowing for three missed cleavages at lysines and two at arginines. Precursor precision was set at 10 ppm with fragment-ion precision set at 20 ppm.

Electron microscopy. Negative-stain electron microscopy was performed at the Bio21 Advanced Microscopy Facility, the University of Melbourne and Ramaciotti Centre for Cryo-EM, Monash Univesrity. Three microlitres of purified Rh5–CyRPA–Ripr complex was incubated on glow-discharged holey carbon grids (Quantifoil 1.2/1.3) with a 5-nm continuous carbon support layer for 30 s. Excess sample was removed by blotting on a filter paper, and grids were washed in water before staining in 1% uranyl acetate solution for 30 s. Grids were air-dried and transferred to a FEI TF30 electron microscope operated at 200 kV, with images recorded at a calibrated magnification of 20,500 at defocus values that ranged from 1 to 2 μm.

For cryo-EM, frozen samples were transported on dry ice to Janelia CryoEM facility. Before grid preparation, an aliquot of protein was thawed on ice, immediately followed by glycerol removal using a 0.5-ml 100k mwco Amicon filtration unit (Millipore) in a 4-°C table-top centrifuge at 2,000 rcf for minimum of 5 cycles. Then, 3.2 μl of sample diluted in glycerol-free buffer was applied to a glow-discharged 200-mech quantifoil 1.2/1.3 Au grid (Quantifoil), and rapidly plunge-frozen into a liquid ethane bath on a Vitrobot (FEI).

Grids were imaged on a 300 kV FEI Titan Krios cryo-electron microscope (FEI) equipped with a spherical aberration corrector, an energy filter (Gatan GIF Quantum) and a post-GIF Gatan K2 Summit direct electron detector. Images were taken on the K2 camera in dose-fractionation mode at a calibrated magnification of 48,077, corresponding to 1.04 Å per physical pixel (0.52 Å per super-resolution pixel). The dose rate on the specimen was set to be 9.25 electrons per Å^2 per second and total exposure time was 10 s, resulting in a total dose of 92.5 electrons per Å^2. With dose fractionation set at 0.2 s per frame, each movie series contained 50 frames and each frame received a dose of 1.85 electrons per Å^2. An energy slit with a width of 20 eV was used during data collection. Fully automated data collection was carried out using SerialEM with a nominal defocus range set from −1.5 to −3 μm.

The first round of data collection and processing indicated that there are a limited number of projection views of the sample. To get more projection views of the sample, four different datasets were collected with the compu-stage tilted at different angles. To be specific, 6,104 movie series were collected at 0 degree tilt; 2,485 movie series were collected at 30 degree tilt; 1,709 movies were collected at 40 degree tilt and 2,676 movies were collected at 45 degree tilt.

Image processing. Beam-induced motion was measured, corrected and dose-weighted at 1.85 electron per Å^2 per frame with data binned by 2 using cisTEM[16]. CTF determination for each movie series was calculated by amplitude averaging of every 3 frames using cisTEM. Automated particle picking using ab inito mode was carried out in cisTEM on all the micrographs and particle stacks were extracted for each dataset. For data collected at 30, 40 and 45 degree tilt, local CTF correction was performed for each particle using GCTF[17]. Multiple rounds of reference-free 2D classification with CTF correction was performed for each dataset in cisTEM to throw away bad particles. Particles from good representative 2D classes from all four datasets were combined to form a new stack of 752,018 particles. This new particle stack was loaded into cryosparc to generate ab initio 3D models[18]. The resultant initial models and heterogeneous refinement in cryosparc indicated that the particles belong to two different populations: 70.9% of particles are CyRPA–Ripr binary complex and 29.1% are Rh5–CyRPA–Ripr ternary complex. The two populations were separately imported into cisTEM for further 3D refinement. Fourier shell correlation (FSC) at a criteria of 0.143 reported a resolution of 5.07 Å for the binary complex and 7.17 Å for the ternary complex.

Model building and refinement. The crystal structures of Rh5 (RCSB Protein Data Bank code (PDB ID): 4WAT)[5] and CyRPA (PDB ID: 5TIK)[9] were individually docked into the Rh5–CyRPA–Ripr ternary map using UCSF Chimera [19] The fitted Rh5 and CyRPA models were manually refined in Coot[20]. Because densities corresponding to the N-terminal β-hairpin and part of the C-terminal tail of Rh5 were disordered, these domains were removed from the model. The N-terminal α-helix and β-strand of Ripr that contact blade 6 of CyRPA were manually built as a poly-alanine model, guilded by the density map in Coot. After manual building in Coot, the model of Rh5–CyRPA–Ripr was globally real-space-refined and minimized in Phenix[21] using the Rh5–CyRPA–Ripr density map. During the course of manual model building and global refinement in phenix, torsion, rotamer, Ramachandran, C-β deviation restraints and secondary structure restraints were applied throughout. After model refinement, Bsoft package was used to calculate FSC curves between refined atomic models and density maps[22]. All structural figures were generated in UCSF Chimera[19] and pymol (www.pymol.org).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

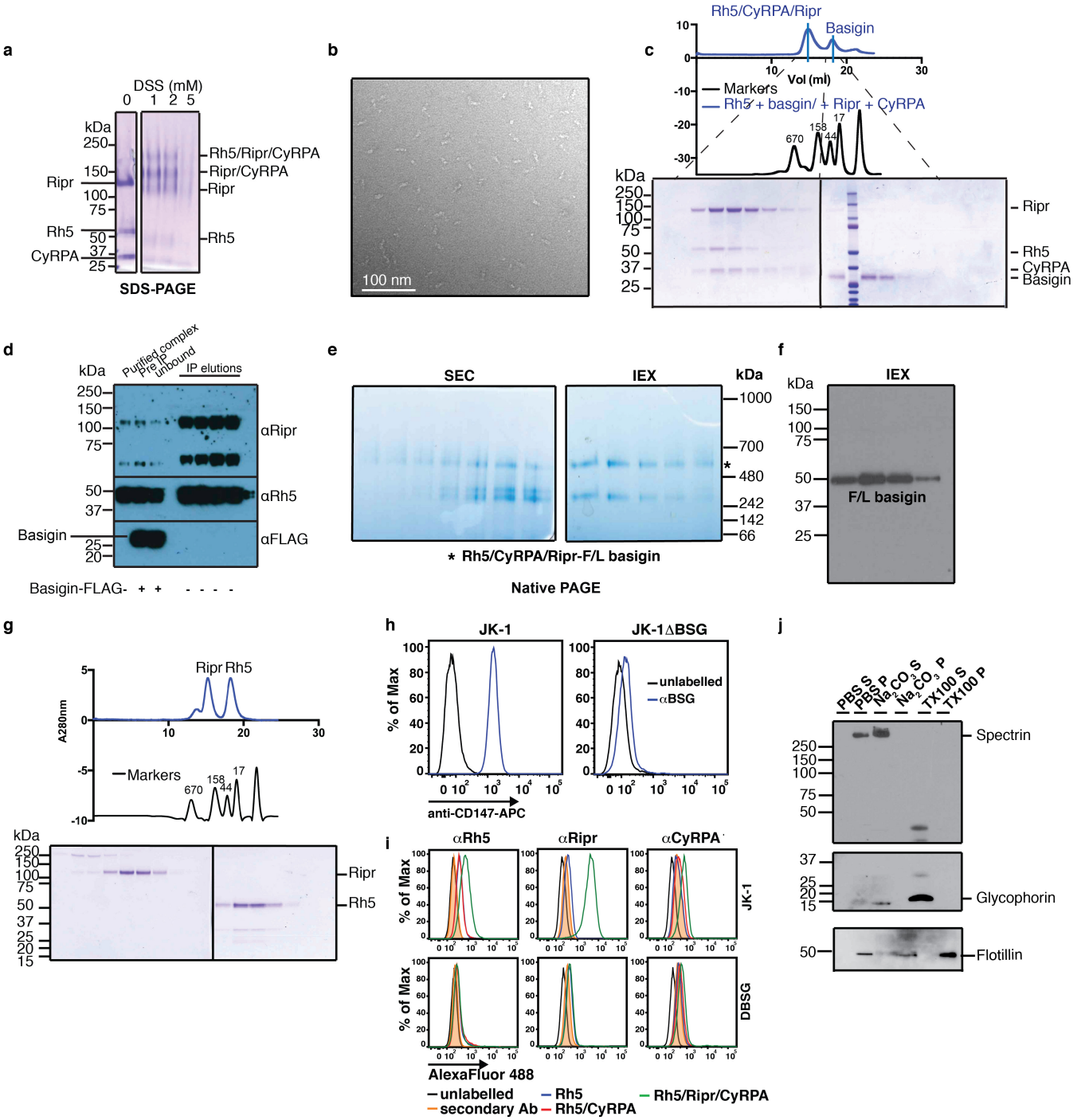
All relevant data are available from the authors and/or are included with this Letter. Atomic coordinates and the cryo-EM density maps have been deposited in the PDB under accession number 6MPV and the Electron Microscopy Data Bank under accession numbers EMD-9192 and EMD-9193.

13. Douglas, A. D. et al. A PfRH5-based vaccine is efficacious against heterologous strain blood-stage *Plasmodium falciparum* infection in aotus monkeys. *Cell Host Microbe* **17**, 130–139 (2015).
14. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
15. Cox, J. et al. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
16. Grant, T., Rohou, A. & Grigorieff, N. *cisTEM, user-friendly software for single-particle image processing. eLife* **7**, e35383 (2018).
17. Zhang, K. GCTF: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
18. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).

© 2019 Springer Nature Limited. All rights reserved.

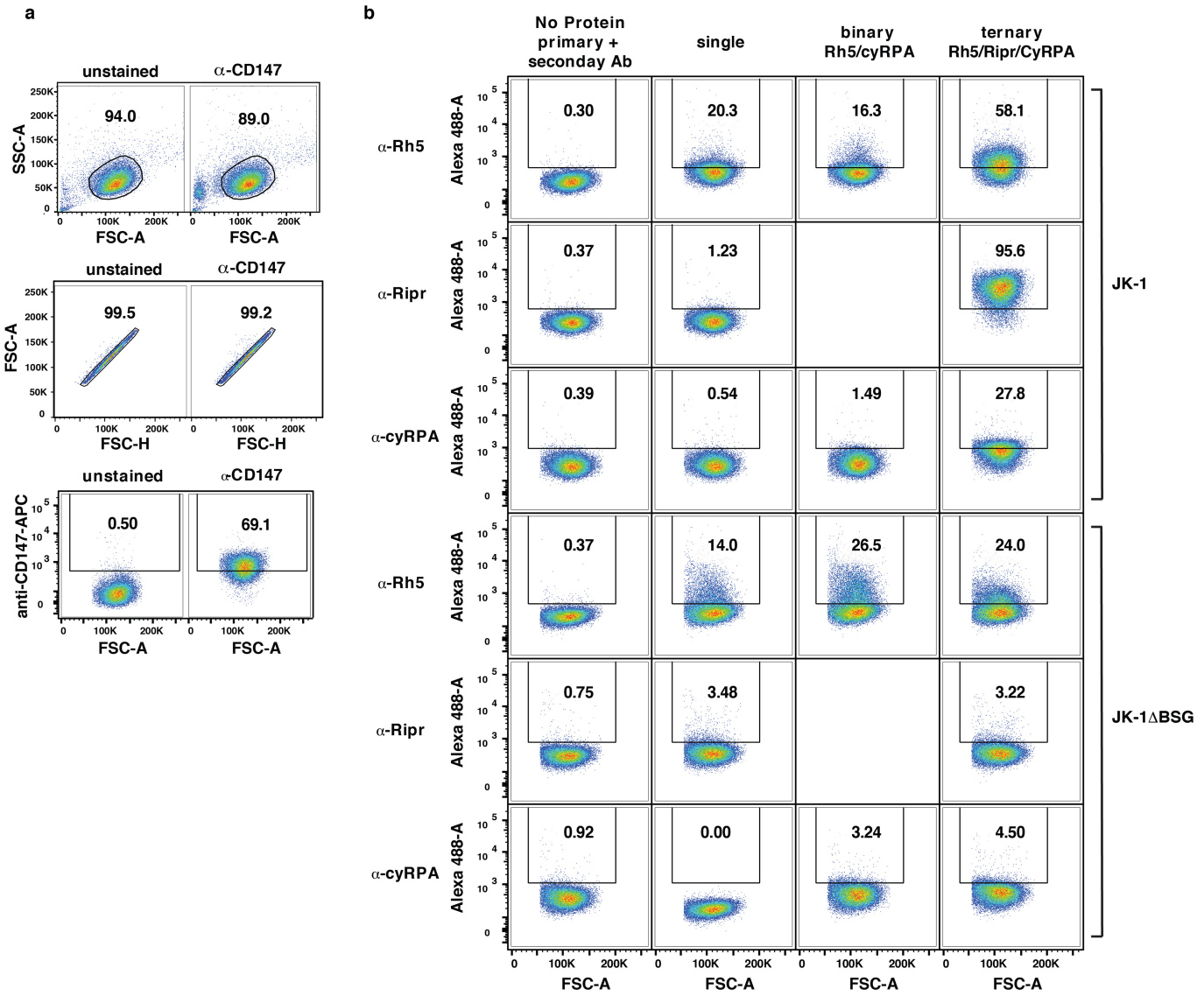
19. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
20. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
21. Adams, P. D. et al. Ensemble refinement, CABLAM. *Comput. Crystallogr. Newsl.* **4**, 28–32 (2013).
22. Heymann, J. B. & Belnap, D. M. Bsoft: image processing and molecular modeling for electron microscopy. *J. Struct. Biol.* **157**, 3–18 (2007).
23. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protocols* **10**, 845–858 (2015).

© 2019 Springer Nature Limited. All rights reserved.



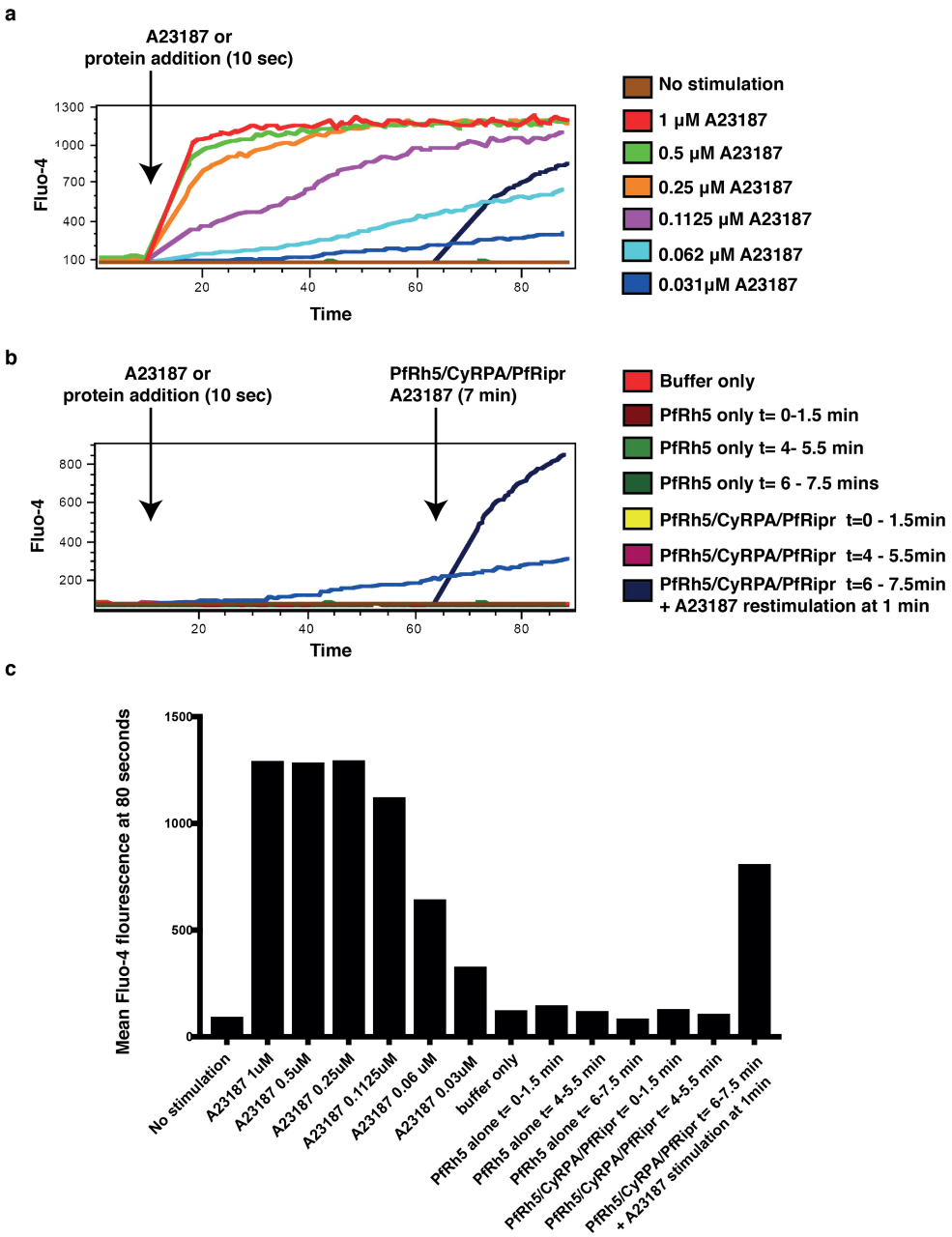
Extended Data Fig. 1 | Stoichiometry of the Rh5–CyRPA–Ripr complex, interaction with soluble or full-length basigin and JK-1 cells. **a**, Chemical cross-linking of Rh5–CyRPA–Ripr complex by disuccinimidyl suberate (DSS), analysed on an SDS–PAGE. **b**, Negative-stain electron microscopy of purified Rh5–CyRPA–Ripr ternary complex, showing the elongated shape of the complex. **c**, Size-exclusion chromotography analysis of a mixture containing recombinant Rh5–CyRPA–Ripr complex and soluble basigin. Soluble basigin was eluted separately from the ternary complex, indicating no binding to basigin. **d**, Immuno-precipitation of Rh5–CyRPA (tagged with haemagglutinin (HA))–Ripr from parasite schizont-stage extract using anti-HA resins could not pull down soluble basigin. **e**, Native PAGE analysis of the size-exclusion chromotography (left) and anion-exchange chromatography (right) eluted fractions, showing the migration of the quaternary complex comprised Rh5–CyRPA–Ripr–full-length basigin, close to ~ 600 kDa. **f**, Western blot analysis of the ion-exchange chromatography elutions indicated the presence of full-length basigin in the complex. **g**, Size-exclusion chromotography analysis showed that Rh5 and Ripr are eluted separately from each other, which indicates that no complex is formed in the absence of CyRPA. **h**, Labelling of JK-1 and JK-1ΔBSG cells with anti-CD147 (basigin) and analysis using flow cytometry. **i**, Analysis of Rh5 (blue line), Rh5–CyRPA (red line) and Rh5–CyRPA–Ripr (green line) binding to JK-1 and JK-1ΔBSG (ΔBSG) cells. **j**, Differential solubilization showing that peripheral membrane protein (spectrin), integral membrane protein (glycophorin) and detergent-resistant membrane protein flotillin were localized in the sodium-bicarbonate-soluble, TX100-soluble and TX100-insoluble fractions, respectively. Experiments in **a**–**j** were repeated three times with biologically independent samples, and were reproducible.

© 2019 Springer Nature Limited. All rights reserved.



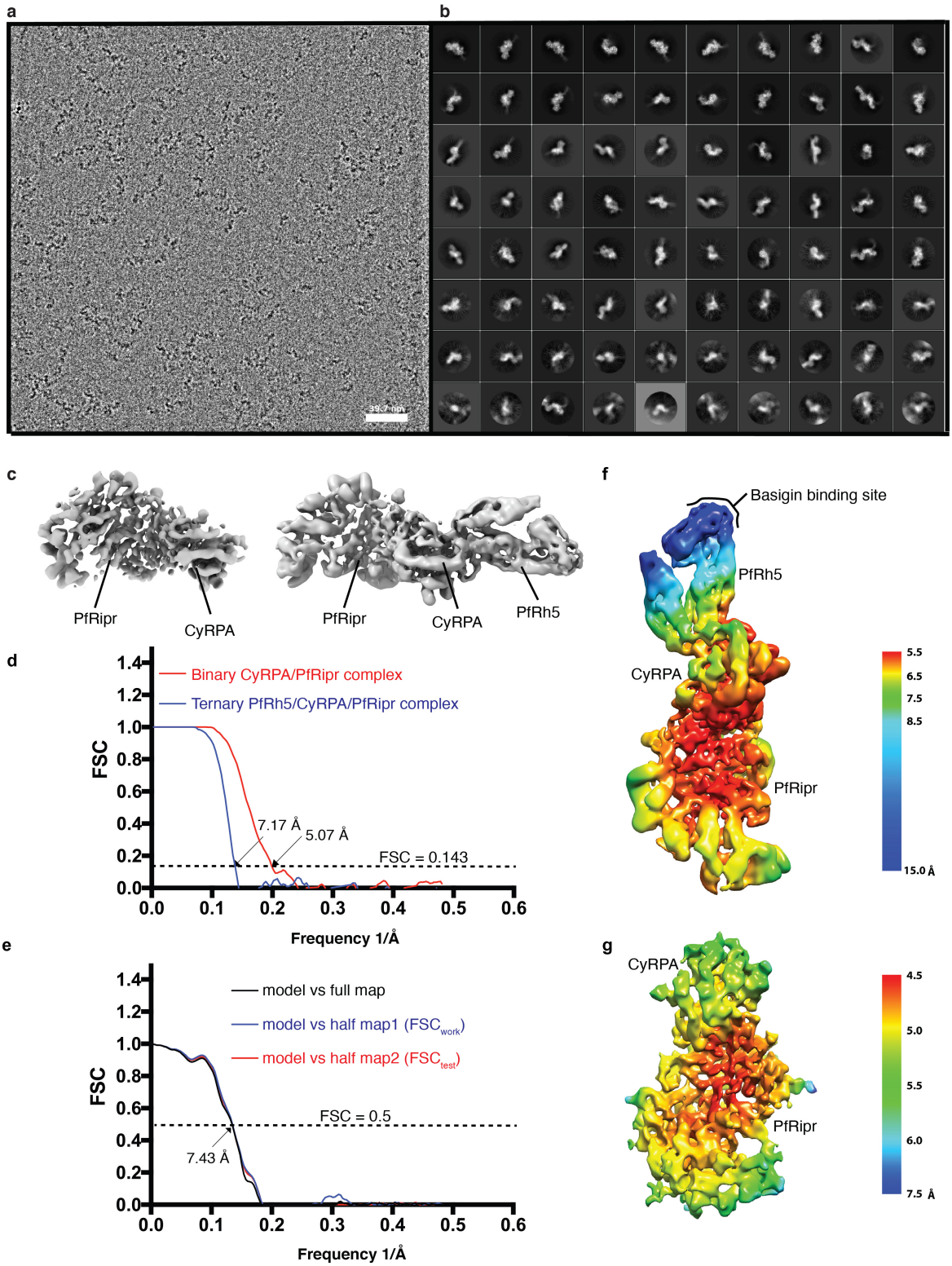
Extended Data Fig. 2 | Gating strategy for FACS analysis. **a**, Representative flow cytometry plots of unstained JK-1 cells and cells stained with anti-CD147-APC. Populations of cells were gated using forward and side scatter (top). Doublet exclusion was performed using FSC-A and FSC-H (middle). The voltage used for the APC channel (anti-CD147) was set using unstained cells, where the negative population was positioned $< 10^3$ (bottom). Experiments were repeated three times independently and were reproducible. **b**, Representative flow cytometry plots of JK-1 and JK-1ΔBSG cells in the presence or absence of recombinant proteins. Cells were incubated with no protein, single recombinant protein, the binary Rh5–CyRPA complex or the ternary Rh5–CyRPA–Ripr complex, followed by the subsequent incubation of the respective primary and secondary antibodies as indicated. Experiments were repeated three times with biologically independent samples, and were reproducible.

© 2019 Springer Nature Limited. All rights reserved.



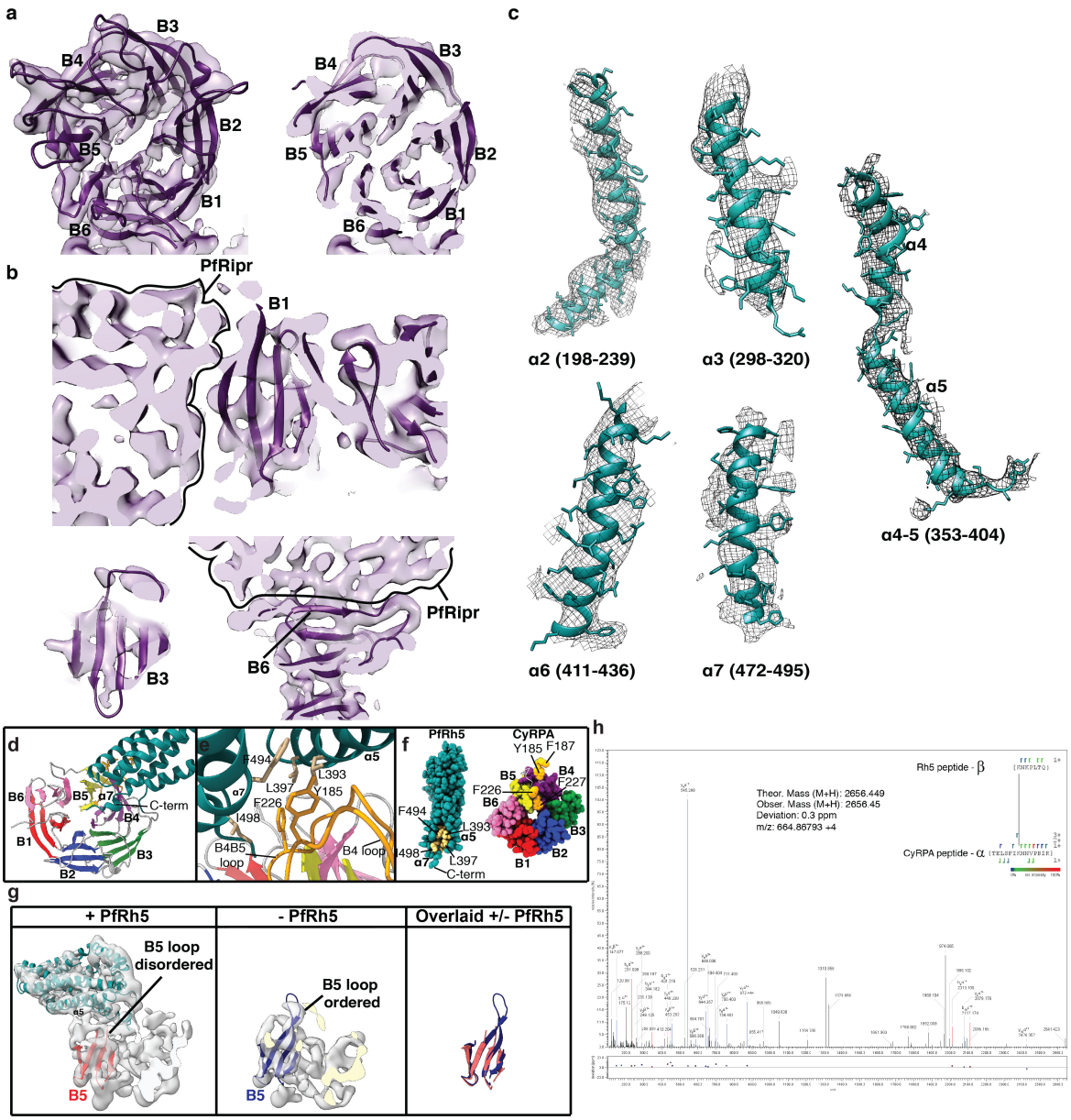
Extended Data Fig. 3 | The Rh5–CyRPA–Ripr complex does not stimulate Ca^{2+} flux across the erythrocyte membrane. a, FACS kinetic plot of online stimulation of Fluo-4-loaded erythrocytes, stimulated with a dilution series of the Ca^{2+} ionophore A23187 (1 μM–0.031 μM). An equivolume of a twofold concentration of A23187 was added to erythrocytes loaded with Fluo-4 at 10 s, and fluorescence was monitored continuously for 1 min 30 s. An equivolume of a 2× concentration of Rh5 only or preassembled Rh5–CyRPA–Ripr complex was also added at 10 s after the start of acquisition. Rh5- and Rh5–CyRPA–Ripr complex-bound erythrocytes were then re-measured at 4 min and 6 min. At 7 min post-addition of Rh5–CyRPA–Ripr, Fluo-4-loaded erythrocytes were re-challenged with 1 μM of A23187. **b**, Kinetic plot of samples to which Rh5 alone or Rh5–CyRPA–Ripr complex was added, plotted only with stimulation with 0.031 μM A23187. **c**, Representation of mean fluorescence intensity values at 80 s, as above. Experiments were repeated three times with biologically independent samples, and were reproducible.

© 2019 Springer Nature Limited. All rights reserved.



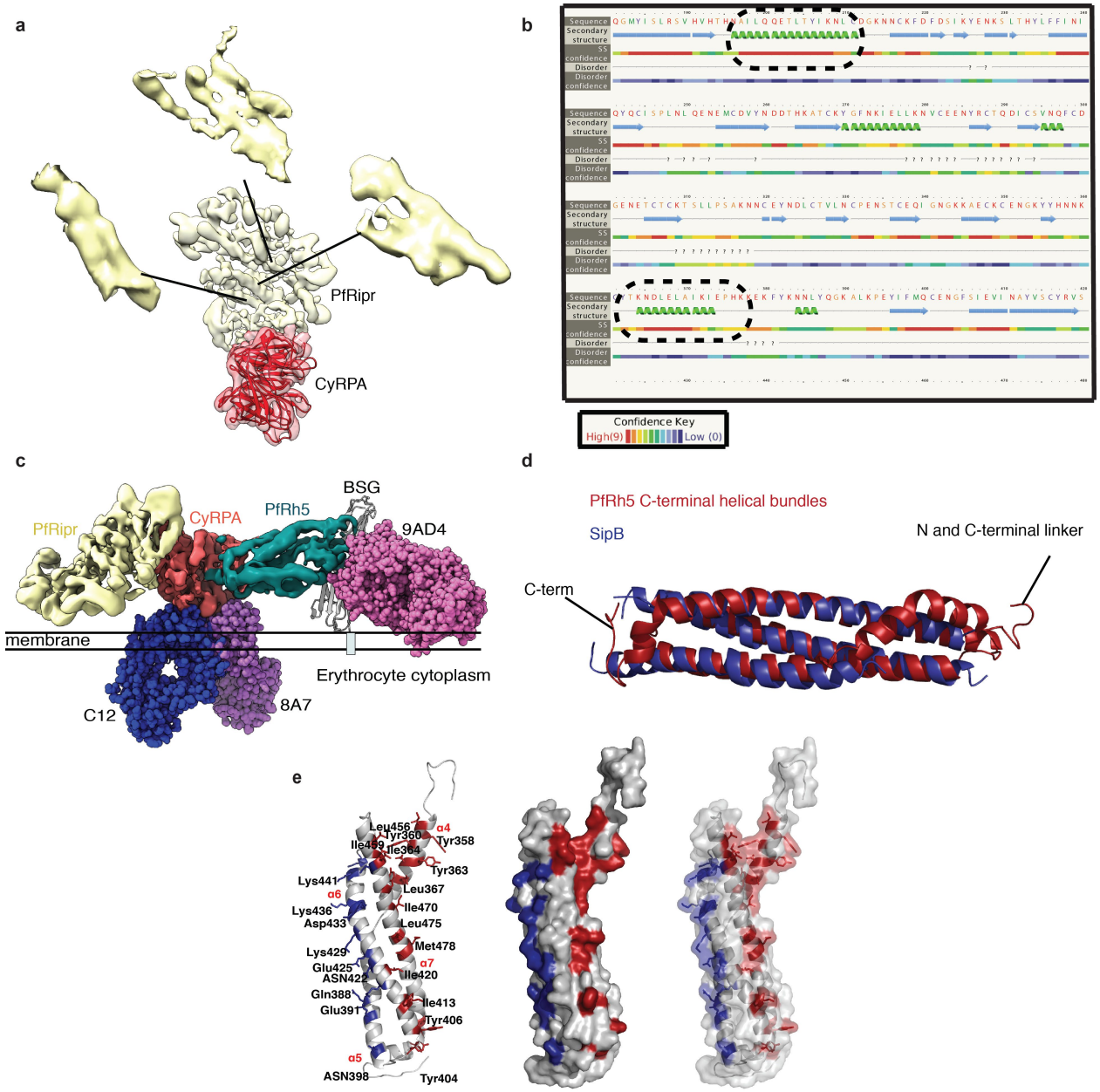
Extended Data Fig. 4 | Cryo-EM single-particle analysis of Rh5–CyRPA–Ripr complex a, A micrograph, after drift correction and dose-weighting. **b**, Reference-free 2D class averages. **c**, 3D classification resulted in separation of the binary CyRPA–Ripr complex (left) and the ternary Rh5–CyRPA–Ripr complex (right). **d**, FSC curves indicating the overall resolutions of the ternary Rh5–CyRPA–Ripr (blue) and binary CyRPA–Ripr (red) reconstructions. **e**, FSC curves between the final refined Rh5–CyRPA–Ripr ternary model and full map, excluding an unbuilt region of Ripr density (black); between the model refined in half map 1 and the reconstruction from that same half (FSC_{work}, blue); and between the model refined in half map 1 and the reconstruction from half map 2 (FSC_{test}, red) for the Rh5–CyRPA–Ripr ternary complex. **f**, **g**, Local-resolution estimation colour spectrum of the ternary Rh5–CyRPA–Ripr map (**f**) and the binary CyRPA–Ripr map (**g**).

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 5 | Cryo-EM densities of CyRPA in the binary complex and Rh5 in the ternary complex a, Electron microscopy density showing the top view of the CyRPA β-propeller (left), and a cross-section of the same region showing the resolution of the 6-bladed β-sheets of CyRPA (right). **b**, Density of β-strands resolved in blades 1, 3 and 6 of the CyRPA β-propeller. **c**, Electron microscopy densities showing the individual α-helices (α2–α7) of Rh5. **d**, Model showing α7 helix of Rh5 inserted into the central cavity of CyRPA. **e**, Hydrophobic residues (L393, L397, F494 and I498) form a groove of Rh5, in contact with aromatic residues (Y185, F187 and F226) presented by B4 and B4–B5 loops of CyRPA. **f**, Models showing the hydrophobic groove of Rh5 and the binding aromatic residues of CyRPA (Y185, F187 and F226), shown in orange. **g**, Density maps and the refined atomic models showing the disordered B5 loop in CyRPA upon binding of Rh5 (left), the corresponding region showing the ordered B5 loop in CyRPA in the absence of Rh5 (middle) and the two superimposed blade 5 β-sheets of CyRPA in the absence (blue) and presence (red) of Rh5 (right). **h**, Tandem mass spectra of DSS cross-linked peptides identified from tryptic digestion of gel-purified Rh5–CyRPA–Ripr complex. High-resolution spectra from Q-Exactive mass spectrometer for two cross-linked peptides between Rh5(520–526) and CyRPA(37–50). Experiments were repeated three times with biologically independent samples, and were reproducible.

© 2019 Springer Nature Limited. All rights reserved.



Extended Data Fig. 6 | Electron microscopy density map of Ripr in the ternary complex, and orientation of the Rh5–CyRPA–Ripr complex on the erythrocyte membrane. a, Density map corresponding to Ripr (yellow) and CyRPA (red), showing several β-sheets in Ripr. **b**, Secondary structure prediction using the Phyre2 server indicated that two putative high-confidence α-helices (in dashed rectangles) reside in the N terminus of Ripr[23]. **c**, The cryo-EM structure of Rh5–CyRPA–Ripr overlaid with the crystal structure of the Rh5–basigin (BSG) complex. The overlaid structures suggest the Rh5–CyRPA–Ripr complex is positioned parallel to the erythrocyte membrane before insertion. The crystal structures of CyRPA–C12, CyRPA–8A7 and Rh5–9AD4 antigen–antibodies complexes were also overlaid with the Rh5–CyRPA–Ripr cryo-EM structure. The overlaid structures suggest these monoclonal antibodies function to inhibit the docking of the invasion complex to the erythrocyte membrane. **d**, The crystal structure of the N-terminal domain of SipB is superimposed with the C-terminal helical bundles of Rh5. Over 144 residues of SipB were aligned with Rh5 C terminus, with a root mean square deviation of 3.4 Å. **e**, The Rh5 C-terminal helical bundle containing the α4–α7 helices is shown in cartoon and surface representation. Hydrophobic residues lining one side of the helical bundle are coloured in red, whereas hydrophilic residues lining the opposite side of the helical bundle are coloured in blue.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 1 | Mean on- and off-rates and dissociation constants derived from fitting SPR sensorgrams

	k_{a1} ($M^{-1}S^{-1}$)	k_{a2} ($M^{-1}S^{-1}$)	k_{d1} (S^{-1})	k_{d2} (S^{-1})	k_{D1} (M)	k_{D2} (M)
PfRh5	3.62 (± 0.08) x10^{5}	2.30 (± 0.01) x10^{4}	7.91 (± 0.37) x10^{-2}	1.21 (± 0.03) x10^{-3}	2.19 (± 0.15) x10^{-7}	5.27 (± 0.14) x10^{-8}
PfRh5/CyRPA	2.59 (± 0.03) x10^{5}	1.55 (± 0.01) x10^{4}	6.24 (± 0.3) x10^{-2}	3.17(± 0.16) x10^{-3}	2.41 (± 0.14) x10^{-7}	2.04 (± 0.11) x10^{-7}
PfRh5/CyRPA/ PfRipr	2.77 (± 0.08) x10^{5}	1.83 (± 0.05) x10^{4}	6.27 (± 0.49) x10^{-2}	1.12 (± 0.17) x10^{-3}	2.27 (± 0.25) x10^{-7}	6.13 (± 0.11) x10^{-8}

Data are mean values from two independent experiments, with standard error of the mean indicated.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 2 | Kinetic constants derived from fitting SPR sensograms

	k_{a1} ($M^{-1}S^{-1}$)	k_{a2} ($M^{-1}S^{-1}$)	k_{d1} (S^{-1})	k_{d2} (S^{-1})
PfRh5	3.70 (± 0.03) x10^{5}	2.30 (± 0.01) x10^{4}	7.55 (± 0.05) x10^{-2}	1.18 (± 0.01) x10^{-3}
PfRh5	3.54 (± 0.02) x10^{5}	2.29 (± 0.01) x10^{4}	8.28 (± 0.05) x10^{-2}	1.24 (± 0.01) x10^{-3}
PfRh5/CyRPA	2.62 (± 0.01) x10^{5}	1.55 (± 0.01) x10^{4}	5.94 (± 0.03) x10^{-2}	3.01 (± 0.01) x10^{-3}
PfRh5/CyRPA	2.56 (± 0.01) x10^{5}	1.56 (± 0.01) x10^{4}	6.54 (± 0.03) x10^{-2}	3.34 (± 0.02) x10^{-3}
PfRh5/CyRPA/ PfRipr	2.86 (± 0.02) x10^{5}	1.88 (± 0.01) x10^{4}	5.78 (± 0.03) x10^{-2}	0.95 (± 0.01) x10^{-3}
PfRh5/CyRPA/ PfRipr	2.68 (± 0.02) x10^{5}	1.78 (± 0.01) x10^{4}	6.76 (± 0.04) x10^{-2}	1.28 (± 0.01) x10^{-3}

	k_{D1} (M)	k_{D2} (M)	R_{max1} (RU)	R_{max2} (RU)	χ^2 (RU^2)
PfRh5	2.04 (± 0.03) x10^{-7}	5.13 (± 0.07) x10^{-8}	854 (± 3)	133 (± 0.5)	120
PfRh5	2.34 (± 0.03) x10^{-7}	5.41 (± 0.07) x10^{-8}	474 (± 2)	54 (± 0.3)	27
PfRh5/CyRPA	2.26 (± 0.02) x10^{-7}	1.94 (± 0.02) x10^{-7}	726 (± 2)	166 (± 0.3)	56
PfRh5/CyRPA	2.55 (± 0.03) x10^{-7}	2.14 (± 0.03) x10^{-7}	406 (± 1)	74 (± 0.2)	17
PfRh5/CyRPA/ PfRipr	2.02 (± 0.02) x10^{-7}	5.05 (± 0.08) x10^{-7}	582 (± 0.5)	294 (± 0.3)	32
PfRh5/CyRPA/ PfRipr	2.52 (± 0.03) x10^{-7}	7.22 (± 0.10) x10^{-8}	340 (± 0.4)	118 (± 0.1)	9

Data are mean values from two independent experiments, with standard error of the mean indicated.

© 2019 Springer Nature Limited. All rights reserved.

Extended Data Table 3 | Cryo-EM data collection, refinement and validation statistics

Data Collection		
Microscope	Titan Krios	
Acceleration voltage (kV)	300	
Detector	K2 Summit with Quantum Image Filter	
Spherical aberration (mm)	0.01	
Filter slit width (eV)	20	
Defocus range (μm)	-1.5 to -3	
Number of micrographs	12,974	
Magnification - nominal	105,000	
Magnification - calibrated	48,077	
Pixel size (Å/pixel)	1.04	
3D reconstructions	PfRh5/CyRPA/PfRipr	CyRPA/PfRipr
Number of particles	218,837	533,180
Resolution (Å)	7.17	5.07
Atomic model	**PfRh5/CyRPA/PfRipr**	
Non-hydrogen atoms	5091	
Total number of protein residues	612	
Residues in PfRh5	276	
Residues in CyRPA	315	
Residues in PfRipr	21	
r.m.s. deviations		
Bond lengths (Å)	0.006	
Bond angles(°)	1.09	
Validation		
Molprobity score	2.23	
Clashscore, all atoms	15.14	
Rotamer ouliers (%)	0.18	
Ramachandran plot		
Favoured	90.1	
Allowed	9.9	
Outliers	0	

© 2019 Springer Nature Limited. All rights reserved.

TOOLBOX

3D PRINTING IN THE LAB

As the cost of 3D printers tumbles, researchers have begun using them to make everything from bespoke equipment for experiments to realistic models of human organs.

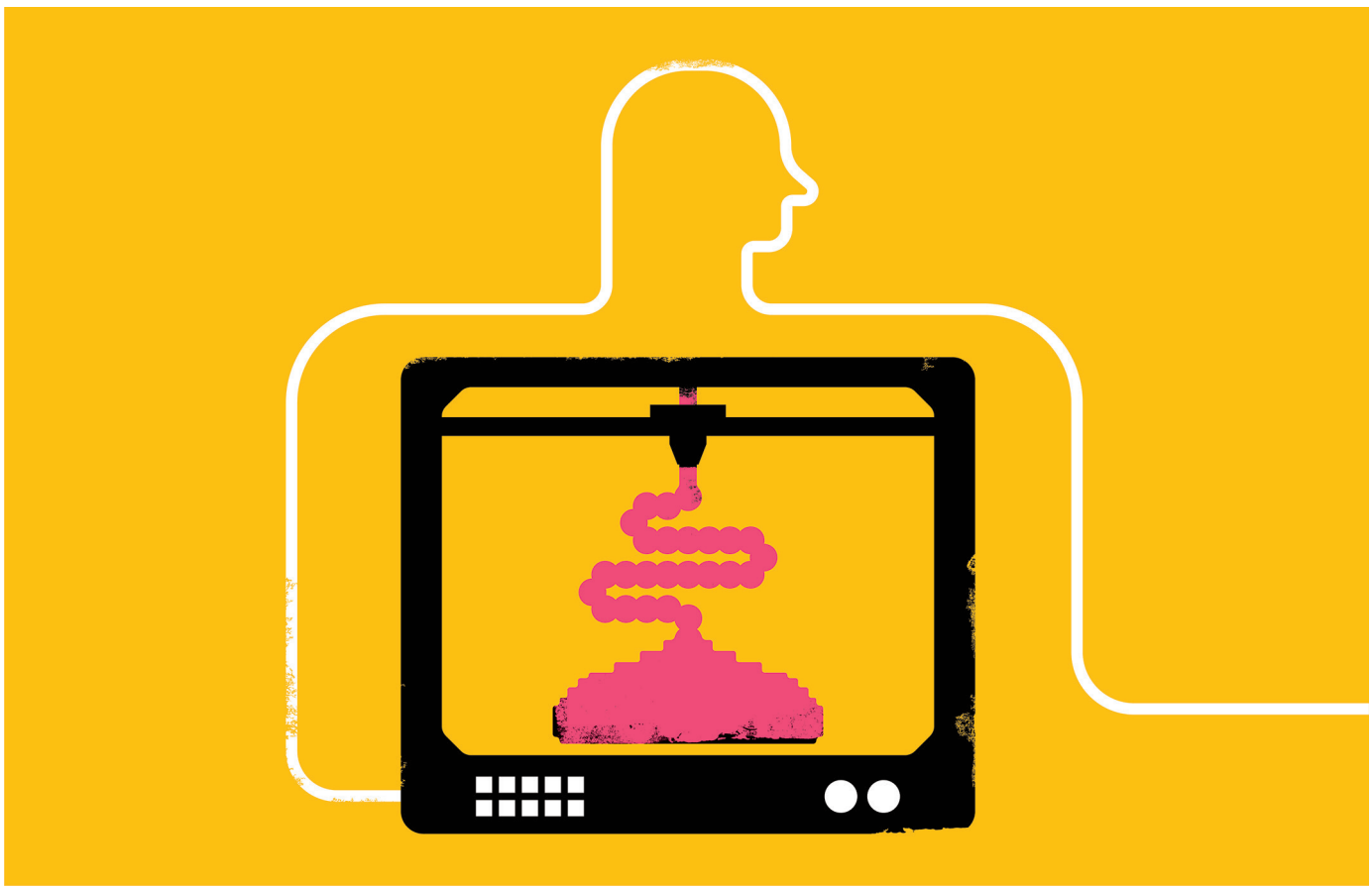


ILLUSTRATION BY THE PROJECT TWINS

BY ANDREW SILVER

Valentine Ananikov, a chemist at the Zelinsky Institute of Organic Chemistry in Moscow, runs chemical reactions so delicate that just a trace of metal nanoparticles, smaller than a bacterium, could change his results. So when his laboratory finishes an experiment, rigorous cleaning is required. Or at least, it used to be. In 2016, Ananikov began creating disposable reaction vessels instead. To do that, he relies on a technology that has captured the imagination of do-it-yourself hackers, engineers and scientists alike: 3D printing.

In 3D printing, also known as additive manufacturing, a 3D computer model is transformed into a physical object layer by layer, like icing a cake. Ananikov's team uses the technology to create bespoke chemical reactors in days, rather than waiting weeks or more for them to be made and shipped by an outside vendor. More importantly, the cost of 3D printing plastic is so low that the group can afford to treat the equipment as consumables to be used once and then thrown away, with no clean-up required. "For research labs dealing with interdisciplinary projects," Ananikov says, "3D printing is a kind of standard tool nowadays."

3D printers have been widely adopted by members of the 'maker culture' for education and creating innovative objects. But they are increasingly becoming standard equipment in scientific laboratories, as well. Researchers can use them to replace broken instrument parts, build custom sample holders and model everything from biological molecules to oil-bearing rocks. And clinicians can use them to create implants and teaching models.

Objects can be 3D printed using several technologies, but one of the most widespread is fused-filament fabrication (FFF), also called fused-deposition modelling. In FFF printers, a narrow, coloured filament — typically plastic wire — is heated and extruded, forming a shape a layer at a time. By contrast, older stereolithography printers use a tank of liquid light-activated resin that is hardened into precise ▶

© 2019 Springer Nature Limited. All rights reserved.

shapes with a laser. FFF printers tend to produce less detailed objects than stereolithography printers, but are easier and cheaper to use.

Commercial FFF printers can be acquired for anything from hundreds to thousands of dollars. Or researchers can build the hardware themselves with kits or designs from the open-source RepRap project for just a few hundred dollars.

3D printing isn't new: stereolithography printers have existed since the 1980s. But falling prices have made the technology widely available. Below are four ways in which researchers have taken advantage of 3D printing.

EQUIPMENT ON THE GO

Julian Stirling, a physicist at the University of Bath, UK, is part of a team that designed light microscopes that can be made with 3D-printed plastic components. The idea is to build them in the field in Tanzania and use them to diagnose malaria by searching for parasites in blood. Tanzania has a shortage of knowledgeable mechanics and local components for repairing scientific equipment, he says, and importing components can be expensive and time-consuming. By 3D printing parts, local doctors and scientists can repair their microscopes more quickly and cheaply. A local firm in Tanzania has even created FFF printers from electronic waste and other local materials, he adds.

Several websites, including Thingiverse and MyMiniFactory, provide forums for scientists to share computer models of printable components. But in Stirling's experience, models on these sites are often incomplete, lacking either documentation for a particular project or key files for modifying the designs. As a result, his team creates its builds from scratch, using an open-source programming language called OpenSCAD. Their microscopes can be entirely 3D printed except for the camera, motors and lenses.

When it comes to 3D printing, it's easy to make mistakes, Stirling says. But because the technology is fast and inexpensive, it's simple to iterate on designs. "This experience can only be built up by trial and error," he notes.

Practice has taught Stirling that there's a big difference between using a 3D printer in the laboratory and doing so in the field. 3D printing plastic filament in Tanzania's humid climate is typically harder than in a climate-controlled laboratory because the humidity affects the plastic filament, leading to more failed prints. Furthermore, power cuts are not uncommon, and only some printers can resume printing a half-finished object after power is restored. There's not much that Stirling and his team can do about the climate, but they do use uninterruptable power supplies to ensure their print jobs run to completion, he says.

LIFE-LIKE ORGANS

Ahmed Ghazi, a urological surgeon at the University of Rochester Medical Center in New York, uses 3D printing to create non-functional human organs, which surgeons can use to practice robot-assisted surgery. For relatively simple procedures, such as removing a spleen, there is little need for such practice. But more complex procedures, such as excising a tumour, can vary wildly from patient to patient. As Ghazi notes, "Tumours are not in textbooks."

Ghazi starts with 3D computer-assisted tomography scans of the patient's tissue, then feeds the data into the commercial medical modelling software Mimics, from Materialise in Leuven, Belgium, and Meshmixer, a free tool from Autodesk in San Rafael, California, to create 3D models. He then prints those models as hollow plastic moulds using an FFF printer, inserts blood-vessel replicas that will connect to a fake-blood pump, and injects the mould with a hydrogel that will solidify into an object with organ-like stiffness. The resulting structures are realistic enough for surgeons to practice their procedures with real-world consequences, including bleeding.

"This is definitely something that's catching on a lot more."

Ghazi says that he and his team use these models for up to four surgery cases a week. In each case, they create two copies of the models and pick the most accurate representation. And they're training other doctors to apply the technology in fields such as heart and liver surgery. "This is definitely something that's catching on a lot more," Ghazi says.

But imperfections remain. The moulds produced by FFF printers often feature tiny ridges and pits, says Ghazi. Such defects are often too small to see with the naked eye, but are plainly visible to the robotic camera, which could affect the surgeon's experience. Ghazi's solution is to spread a layer of room-temperature wax over the inside of the mould, which fills in the ridges and pits, thus smoothing out the final product. "Those little things make a difference," he says.

REPLICA ROCKS

For Mehdi Ostadhassan, a petroleum engineer at the University of North Dakota in Grand Forks, 3D printing provides a tool for optimizing the extraction of oil and gas from rock.

Ostadhassan prints 'rocks' using programs such as OpenSCAD and the commercial 3D computer-aided design software AutoCAD (from Autodesk) in combination with various 3D printers and materials. These rock models have realistic physical properties, including tiny, detailed pores, and Ostadhassan puts them under physical stress to better understand how liquid flows through their real-life equivalents.

To create the most realistic rocks, Ostadhassan uses a range of printing approaches, including binder-jet technology, in which a liquid binding agent is applied layer by layer to gypsum powder or silica sand. The process produces objects with mechanical properties that closely mimic those of real rocks. But unbound powder can also get stuck in the pores, Ostadhassan says, diminishing the quality of the final product. And for some experiments, he needs to apply a water-repelling treatment to get the 'wettability' right. Stereolithography printers are better at printing rocks with detailed pores to enable the study of liquid-flow properties, but the models they produce are not as strong as binder-jet-printed rocks.

As such, Ostadhassan is collaborating with other researchers to develop a custom printer that can mimic those pores and cracks but still produce models with the same mechanical strength as real rocks.

HEAVY METAL

Today's 3D printers can output a range of materials — but not all of them. "The material for 3D printing is very, very limited," says Yang Yang, chief executive of UniMaker in Shenzhen, China, which makes 3D printers for scientific use. But research in the space is intense, and change is coming. One hot growth area is bioprinting, for use in creating structured biological materials. Jin-Ye Wang, a biomedical scientist at Shanghai Jiao Tong University in China, says that her institution has acquired one such device for use in the classroom. These bioprinters blend cells and hydrogels to create structures such as bones and tumour models.

Another growth area, Yang says, is metals. Metal-capable printers use a beam of electrons or a laser to melt metal powders in defined patterns. Jeremy Bourhill, a physicist at the University of Western Australia in Perth who researches dark matter, is studying the use of laser-based 3D metal printers to build a mesh of superconducting niobium. This could be used to block strong magnetic fields that would interfere with dark-matter detection, Bourhill says.

Using conventional machining to create the mesh would require toxic lubricants and waste a substantial amount of niobium, which is expensive. So Bourhill's team is using high-powered lasers to melt and fuse cross-sections of metal powder together. But because the melting point of niobium is about 2,500 °C, the process requires considerable amounts of power. "Niobium's a really tough material," Bourhill says.

Once upon a time, researchers such as Bourhill would have been limited in their options. But with the increased availability of 3D printers, a fundamental shift has occurred, says Yusheng Shi, a materials engineer at the Huazhong University of Science and Technology in Wuhan, China: 3D printing is enabling personalized manufacturing, supplanting centralized manufacturing. As these examples show, researchers have just scratched the surface of what they can do with that power. ■

Andrew Silver *is a science writer based in Taipei.*

© 2019 Springer Nature Limited. All rights reserved.

ONLINE Career resources from our scientific community **nature.com/careers**

MENTAL HEALTH An online resource for advice and support **go.nature.com/wellbeing**

SOCIAL Follow us on Twitter **twitter.com/naturejobs**

GETTY



Many postdoctoral researchers find it hard to compete for non-academic jobs, two studies reveal.

TRAINING

The perils of a postdoc job

Skills learnt as a postdoctoral researcher might not advance your career outside the lab.

BY CHRIS WOOLSTON

Postdoctoral-research stints might not offer the smooth path into a scientific career that many junior researchers are hoping for. The job might even leave them ill-prepared for their professional futures, according to two studies that explore the realities of postdoc life at major research institutions in the United States and Europe.

One study explored the 'mismatch' between the skills sought by employers and the skills learnt in postdoctoral positions at five institutions, including four leading US universities (C. S. Hayter and M. A. Parker *Res. Pol.* http://doi.org/cw62; 2018). Another investigated the postdoc recruitment and hiring procedure at four European universities — a process, the authors argue, that undermines junior scientists' long-term employability and job security (C. Herschberg *et al. Scand. J. Mgmt* **34**, 303–310; 2018).

Postdocs often aspire to gaining tenure-track positions at universities, and the 97 researchers interviewed in 2016 and early 2017 at the 5 institutions were no exception: 84 had originally planned to go on to an academic career. Only seven took their postdoc with a specific plan to one day work outside academia. At the time of the study's publication, 5 of the 97 postdocs had landed tenure-track positions, but many of the rest will have to pursue other options, says lead author Christopher Hayter, a higher-education researcher at Arizona State University in Tempe. "That's par for the course," he says.

For a fuller picture of the employment outlook for postdocs, the study included interviews with 9 principal investigators (PIs) and 16 industry representatives. In general, the PIs interviewed for the study showed little interest in helping their postdocs to prepare for a future career, especially if that preparation took time away from the postdoc's research project.

Hayter and his co-author, management researcher Marla Parker at California State University in Los Angeles, note that the PIs were not actively trying to stall or sabotage postdocs' careers. The authors point out that PIs aren't in the career-counselling business, and that many are not familiar with the concept. "I never had someone hold my hand ▶

© 2019 Springer Nature Limited. All rights reserved.

MEDIA

Film falsities

The portrayal of jobs in science, technology, engineering and mathematics (STEM) in US entertainment media offers mixed messages to girls and women, finds an analysis. The report comes from the Geena Davis Institute on Gender in Media, a non-profit organization supported by Mount Saint Mary's University in Los Angeles, California, and the Lyda Hill Foundation, a private science funder in Dallas, Texas. The longitudinal study examined STEM characters in film and television, and found that the roles largely reinforce the narrative that scientists are white men (see go.nature.com/2qrw4vc). Male STEM characters outnumber female ones by 62.9% to 37.1%, and 71.2% of STEM characters are white. The study also found that films and television shows perpetuate the myth that the physical sciences and engineering, among other disciplines, are inappropriate for women. Female STEM characters are, however, as likely as male ones to be portrayed as leaders in their fields, and are shown as equally competent as, and more intelligent than, men in these roles.

AUTHORSHIP

Who reads whom

Scientific studies published by female authors across 100 topics attract between 2% and 6% more undergraduate student readers in the United States, the United Kingdom, Turkey and Spain than do articles by male authors, according to a study. Using Mendeley, a computer program that manages and shares research papers, the author collected reader data from these four countries plus India in 2014 for articles in 100 subject categories (M. Thelwall *J. Altmetr.* **1**, 3; 2018). He calculated the mean number of readers by gender, field, occupation and position — whether the reader was a student or a senior faculty member. The findings suggest that female authors might have an unrecognized effect on students' education. The author cautions early-career scientists, particularly female researchers, to look beyond citations for evidence that their research might have a broader impact than that metric alone indicates.



through my PhD or postdoc," one PI said in the study. "I had to figure things out on my own, so I now expect the same of my postdocs."

Most PIs did show some willingness to write letters of recommendation or to make phone calls to help their postdocs land tenure-track positions at universities, partly because the success of their trainees in academic posts would burnish their own reputations. But many PIs in the study had little incentive to help trainees to find jobs in industry.

The postdocs who were surveyed also commented on the perceived lack of support from supervisors. "I learnt that you're doomed if you think your PI is going to provide career guidance," said one. Another put it more bluntly: "A lot of PIs just love to squash any interest you have, other than being chained to the bench." Some found ways around the resistance. "If I need to go to a [non-academic career] workshop, I just lie," a postdoc said.

Hayter and Parker's interviews with industry representatives underscored another problem: postdocs can have a hard time competing for non-academic jobs. One potential employer said that postdocs "have all the academic science skills you don't need, and none of the organizational skills that you do". Industry representatives often reported that PhD and master's students generally have an easier time acclimatizing to non-academic careers than do postdocs.

> ***"I learnt that you're doomed if you think your PI is going to provide career guidance."***

To partly remedy that mismatch, Hayter suggests, more universities could offer programmes that teach postdocs entrepreneurial skills. "They may not decide that they want to be entrepreneurs, but it would at least open their minds to other possibilities," he says. One of the universities in his study did establish an entrepreneurship-support programme that has become an important career-development resource for postdocs. Hayter and Parker note that the programme faced opposition from faculty members who thought that it distracted postdocs from their main jobs.

The entire system for hiring, recruiting and training postdocs isn't necessarily geared towards setting trainees up for success, says Channah Herschberg, lead author of the study in the *Scandinavian Journal of Management* and a PhD student in management at Radboud University in Nijmegen, the Netherlands. The authors' interviews with 21 PIs in Switzerland, the Netherlands, Italy and Belgium suggest that PIs mainly want to hire postdocs who can help the lab in the short term, even if they aren't perfect for the position. One Swiss interviewee said that he generally hires a postdoc "who can start immediately, who will be good for the project, but perhaps not super brilliant, not top class".

Respondents also said that the hiring process is often based on informal connections and familiarity. As one Swiss respondent put it, a phone call from a colleague can carry much more weight than do publications or impact factors when it comes to hiring postdocs. PIs are especially likely to hire a postdoc who has already worked in their lab or at least collaborated on a shared project. "PIs have limited time, so they have a preference for people they already know," Herschberg says. In every case, PIs in the study had total control over whom they hired.

The study also describes how PIs generally hire postdocs to complete specific projects that the PIs have conceived and designed. As a result, Herschberg says, postdocs often lack a sense of ownership of or personal accomplishment with their work. And when the time comes to move on to another job, they might not be able to take credit for original ideas. "A lot of postdocs don't have the opportunity to develop their own line of research," she says.

Herschberg notes that the qualities that PIs look for in a postdoc — including availability, familiarity and a willingness to work on a short-term project — aren't necessarily the qualities that will produce the best science or provide the best preparation for a future career. It would help, she says, if funding agencies could give researchers more time to complete their projects, which could translate to longer contracts for postdocs. This, in turn, would give postdocs more job security and opportunities to develop their own skills and ideas.

More-formal recruitment processes that find the best candidates for a given position would also be a step in the right direction, Herschberg says. "If PIs could advertise more openly for positions, they could give opportunities to new people, and the quality of the research might improve," she says. "We want to recruit the most excellent researchers to our universities, but that doesn't always happen when it comes to postdocs."

Sibby Anderson-Thompkins, director of the Office of Postdoctoral Affairs at the University of North Carolina in Chapel Hill, says that the two studies highlight the precarious employment situation, both present and future, for postdocs in the United States and Europe. "They really homed in on some of the challenges and problems," she says. "We refer to postdocs as trainees, but they aren't getting the opportunity to really train for different careers." She notes that PIs are under pressure to complete projects on time and to win their next grant, so cannot always commit themselves to the future of their trainees. "There needs to be a whole retooling of how we design postdoctoral training and how we recruit and hire people into these positions," she says. ■

Chris Woolston *is a freelance writer in Billings, Montana.*

ADAPTED FROM NADIA BORMOTOVA/GETTY

© 2019 Springer Nature Limited. All rights reserved.

COLD MEMORIES

Time to go home?

BY LAURENCE RAPHAEL BROTHERS



Image: An expanse of brown dirt and grey ash. Caption: *2112. Saskatchewan after the great fire. Two million hectares. Our roses, father's prizewinners, all gone. We moved north after that.*

Video: A walled compound in the midst of farmland, cultivator bots in the distance. A teenager waving at the camera. Sirens sound, the youth turns to run into a bunker, and the camera jerkily follows. Voiceover: *2114. Our estate near Yellowknife. Rice paddies and citrus orchards. False alarm this time. The Greater American Army fell apart before they even got to Edmonton. Out of food. Father was terribly upset. He wanted us to buy places in the Syrtis development on Mars. Mother wouldn't hear of it.*

A life-story in pictures and videos. The prospector was the little kid, and the teenager too. His family were wealthy farmers, who stayed on too long as dollars became worthless and their crops withered and rotted. Of the whole clan, only he was able to secure a spot on a refugee shuttle, way too late to choose a nice destination like Mars or Luna. He wound up one of the first inhabitants of Rock City here in the Neptune Trojans. Mars Council made LongReach exocorp settle a million refugees. They built Rock City as a depot for their bot miners to work the asteroids. Settling refugees was an afterthought.

After the prospector arrived here, there were no more images, no more videos. Everything was text, mostly about how much he hated Rock City. But everyone in Gen-1 hated it; we've only recently got to the point of affording a few luxuries. I almost skipped the most important bit. He hit it rich before I was even born. A 750,000-kilowatt-hour bounty from LongReach for a lode of pure iridium. He'd achieved the Rock City dream. So why was he out here frozen to death? He wanted to go home. To a ruined, blighted planet, but home, nevertheless.

Sol 256, 2142. Took six months for LongReach to answer my enquiry. That's why they planted us so far off. So we can't bother them, so we'll be forgotten. They say it's a million kwuh for a spot on an inner system freighter. And then another 500,000 for a residency at the Spitzbergen climate-monitoring station. The temperature has dropped 0.1 Celsius since 2130: there's hope for the future! The iridium bounty would cover my costs ten times over if it was in marsbucks instead of kwuh. Bastards. Still, one more strike is all I need.

The final entry. I didn't want to read it. But I had to.

Sol 18, 2185. I have to accept it. I'm never going home. The dream is over for me. For you, though — if you're reading this — open the battery case on this tablet. It's all I have left.

I almost didn't do it. Too sensitive, I guess. But I found two items. First, a plastic slip with an account number and passcode. Second, a frail, brittle blossom, just a hint of pink left in it. I looked it up online: it was a dried rose flower. The account turned out to have almost the whole bounty the prospector had earned. Mine, now. The flower — well, it didn't mean much to me at first. But I checked the latest climatology reports. Another half-degree drop in the past 40 years. Maybe the planet is healing. Maybe I'll take that flower back to Earth someday. Maybe I'll see roses blooming there myself. If I strike it rich … ■

Laurence Raphael Brothers *is a writer and a technologist. For more stories, visit his website, https://laurencebrothers.com. You can follow him on Twitter: @lbrothers.*

ILLUSTRATION BY JACEY

NATURE.COM
Follow Futures:
@NatureFutures
go.nature.com/mtoodm

© 2019 Springer Nature Limited. All rights reserved.